UNIVERSITÉ DU QUÉBEC À MONTRÉAL

GOLIATE : UN NOUVEAU TEST D'ASSOCIATION GÉNÉTIQUE COMBINANT LE PROCESSUS DE COALESCENCE ET LES MODÈLES LINÉAIRES MIXTES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN MATHÉMATIQUES

PAR

ABDELHAKIM FERRADJI

AOÛT 2016

REMERCIEMENTS

Je tiens tout d'abord à exprimer toute ma reconnaissance envers mes directeurs de recherche, Mr Oualkacha Karim et Mr Fabrice Larribe.
Merci pour votre disponibilité, pour vos précieux conseils, votre confiance et vos encouragements tout au long de ma maîtrise. Ce fut un plaisir de travailler avec vous durant ces deux dernières années, j'ai beaucoup appris avec vous, merci infiniment.

Je voudrais également remercier mes examinateurs, Sorana Froda et Juli Atherton, tous deux professeurs au département de mathématiques à l'UQAM, pour leurs remarques et commentaires éclairants.

Merci à tous les professeurs qui ont contribué de près ou de loin à ma formation à l'UQAM.

Un grand merci à mes très chers parents, qui n'ont cessé de m'encourager, de croire en moi et qui ont toujours donné le meilleur d'eux même pour me voir réussir.

Merci à toute ma famille et à mes amis qui m'ont soutenu durant la réalisation de ce mémoire.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

RÉSUMÉ

Dans ce mémoire nous présentons un nouveau test d'association génétique qui permet d'analyser simultanément un ensemble de SNPs d'une région chromosomique, tout en tenant compte d'une éventuelle présence de structure de population dans l'échantillon d'étude. Ce test est basé sur un modèle linéaire mixte qui capture l'effet de la structure de population grâce à une nouvelle matrice de similarité, construite en utilisant le processus de coalescence avec recombinaison. Des simulations sont effectuées pour tester et comparer les performances de notre nouvelle approche avec quelques tests proposés dans la littérature. Notre test montre un bon contrôle de l'erreur de type 1 en présence de structure de population contrairement aux autres tests, et semble avoir une puissance comparable à celle des autres méthodes dans le cas des variants génétiques rares.

MOTS-CLÉS : processus de coalescence, cartographie génétique, modèles linéaires mixtes, GWAS, déséquilibre de liaison, variants rares.

INTRODUCTION

L'analyse de la diversité génétique au sein des espèces et en particulier chez l'homme est essentielle à la compréhension des processus d'évolution au niveau de la population et au niveau génomique. Au cours des dernières années, le développement de nouvelles technologies de séquençage et de génotypage a mis à la disposition des chercheurs une quantité astronomique de données génétiques qui leur permettent d'explorer de nouvelles hypothèses scientifiques. Les analyses d'associations génétiques sont devenues une tâche commune entre la génétique humaine et les études des maladies humaines (Hirschhorn *et al.*, 2002 ; Hindorff *et al.*, 2009). Ces analyses étudient la dépendance entre le génotype d'un marqueur génétique et le phénotype. Ce dernier peut par exemple représenter le statut d'une maladie dans une population donnée. Plusieurs tests d'association statistique ont été développés afin d'identifier les gènes responsables de maladies génétiques humaines. Un test d'association peut généralement être effectué à l'aide d'un modèle de régression linéaire simple (Balding, 2006) pour chaque marqueur génotypé. Cette approche simple soulève toutefois quelques problèmes. En effet, ces tests d'association demandent habituellement une procédure d'ajustement des tests multiples telle que la correction Bonferroni afin de garantir un taux global approprié de l'erreur de type 1. Cet ajustement entraîne malheureusement un manque de puissance en raison du seuil de significativé extrêmement bas et difficile à atteindre (de l'ordre de 10^{-7}), typique dans les études d'association sur tout le génome (Wu *et al.*, 2010). De plus, l'analyse d'un seul marqueur à la fois peut être mal adaptée à un contexte de maladies complexes, où de multiples marqueurs interagissent les uns avec les autres pour causer la maladie (Schork, 1997). Toutes ces considérations

ont entraînées l'apparition de nouvelles approches qui permettent d'analyser simultanément un ensemble de marqueurs d'une région chromosomique (Wu *et al.*, 2010; Zhang *et al.*, 2011). Ces analyses "multiples marqueurs" ont montré une meilleure puissance que les analyses "simple marqueur", mais souffrent au même titre que ces derniers d'un taux élevé de l'erreur de type 1, lorsqu'une structure de population est présente dans l'échantillon d'étude. En fait, la structure de population reflète la dépendance ancestrale entre les individus de l'échantillon d'étude et conduit en général à de fausses associations si aucune mesure n'est prise pour tenir compte de cette dépendance ancestrale (Astle et Balding, 2009).

Ce mémoire a pour objectif de développer un nouveau test d'association génétique, que nous avons nommé GoLiATe, permettant de tester l'association entre une région chromosomique et un phénotype d'intérêt. Ce test combine dans un modèle linéaire mixte l'information génétique et l'information généalogique (ou ancestrale) additionnelle obtenue à l'aide du processus de coalescence avec recombinaison, dans le but d'essayer de répondre aux problématiques décrites précédemment, notamment le contrôle de l'effet de la structure de population ainsi que le problème des comparaisons multiples.

Le premier chapitre de ce mémoire est consacré à la présentation des concepts génétiques de base permettant au lecteur non initié de se familiariser avec la terminologie génétique que nous allons manipuler tout au long de cet ouvrage. Par la suite, une introduction à la théorie de la coalescence sera décrite au chapitre II. Le chapitre III sera consacré à la présentation de quelques modèles utilisés en cartographie génétique, dont nous nous sommes inspirés pour développer notre test d'association. Le modèle mathématique ayant servi à la construction du test GoLiATe sera par la suite présenté en détails au chapitre IV. Enfin, le cinquième et dernier chapitre de ce mémoire sera consacré à l'évaluation des performances de notre nouvelle approche à l'aide de données simulées.

CHAPITRE I

CONCEPTS DE BASE EN GÉNÉTIQUE

La génétique est la science qui étudie la transmission de caractères morphologiques et biologiques qui passent de génération en génération. Dans les années 1860, un moine autrichien du nom de Gregor Mendel a présenté une nouvelle théorie de l'hérédité sur la base de son travail expérimental avec des plantes de pois. À cette époque on ignorait tout de la méiose et des chromosomes, mais Mendel croyait en l'existence d'unités héritables qui seront appelées gènes en 1906, par le biologiste danois Wilhem Johannsen. Ainsi, les travaux de Mendel forment le point de départ de la génétique moderne.

Ce chapitre est consacré à l'introduction de différentes notions de génétique nécessaires à la compréhension de ce mémoire. C'est pourquoi, nous invitons le lecteur déjà familier avec ces concepts à passer directement au chapitre suivant. Il est à noter que les informations contenues dans ce chapitre, proviennent de diverses sources, notamment le livre *L'essentiel de la génétique* (Benjamin A. Pierce, 2012) ainsi que divers mémoires (Descary, 2012 ; Dupont, 2013 ; Forest, 2010).

1.1 ADN : centre de l'information génétique

C'est à J. Watson et F. Crick (1953) que l'on doit la découverte de la structure de l'acide désoxyribonucléique (ADN). Ce dernier renferme toutes les informations

nécessaires au fonctionnement de l'organisme. Il est composé de deux brins en forme d'hélice. Chaque brin est constitué de l'enchaînement de petites molécules, appelées nucléotides (A : Adénine, T : Thymine, C : Cythosine et G : Guanine) qui codent l'information (Voir figure 1.1 [1]). Les deux brins sont reliés selon la règle suivante : A est toujours relié à T et C à G. Ainsi, si un fragment de brin contient la séquence ACGTTCAGGT, la séquence complémentaire sur l'autre brin est TGCAAGTCCA.

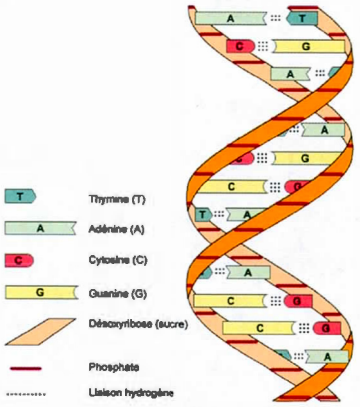


Figure 1.1 Structure en double hélice de l'ADN.

Chaque organisme vivant contient de l'ADN. Chez les eucaryotes [2] et en particulier chez l'homme, l'ADN se trouve dans le noyau de la cellule. En fait, l'ADN est la structure qui compose chaque chromosome présent à l'intérieur du noyau. Chaque cellule humaine contient 46 chromosomes, 23 sont hérités du père et 23 de la mère. Nous avons ainsi 23 paires de chromosomes homologues pour une

1. Source : `http://www.sequencage-genome.com/bases-sequencage-adn`

2. Regroupent tous les organismes, unicellulaires ou pluricellulaires, qui se caractérisent par la présence d'un noyau et de mitochondries dans leurs cellules.

cellule dite diploïde. Si tous les chromosomes étaient déroulés, l'ADN dans une seule cellule formerait un mince fil de près de 2 mètres de long. L'ensemble de ces chromosomes constituent ce que l'on appelle le *génome.*

Un gène est formé par la succession de bases nucléotidique de l'ADN. Le génome humain contient environ 20,000 gènes répartis sur les différents chromosomes et qui codent pour différentes protéines.

Chaque gène possède des versions différentes qu'on nomme *allèles.* L'allèle peut être récessif, dominant ou codominant. Le gène qui contrôle le groupe sanguin (sans tenir compte du rhésus), a par exemple trois allèles différents (système "ABO"). L'allèle "O" est récessif et les allèles "A" et "B" sont dominant. Ainsi, pour un individu qui possède les allèles "O" et "A" sur ses chromosomes homologues, c'est l'allèle dominant qui s'exprimera et l'individu aura pour groupe sanguin "A". Lorsqu'un individu possède le même allèle sur ces deux chromosomes homologues, on dit qu'il est homozygote dans le cas contraire on dira qu'il est hétérozygote.

1.2 Sources de la diversité génétique

On peut se demander pourquoi à l'exception des jumeaux, deux personnes ne sont pas exactement semblables. En fait, la plupart de nos caractéristiques physiques, aussi appelées *phénotypes*, sont le résultat soit d'une mutation qu'a subi notre génome ou d'un processus que nos chromosomes subissent connu sous le nom de recombinaison génétique.

1.2.1 Les mutations génétiques

Une mutation génétique est une modification permanente de la séquence d'ADN, de telle sorte que cette séquence diffère de ce qu'on trouve chez la plupart des gens. Comme une cellule copie son ADN avant de se diviser, il se produit comme une

"faute de frappe" tous les 100,000 nucléotides en moyenne. À chaque fois qu'une de nos cellules se divise, il se produit environ 120,000 fautes[3]. Le plus souvent, une seule base est remplacée par une autre. Parfois, une base est supprimée ou une base supplémentaire est ajoutée. Heureusement, la cellule est capable de réparer la plupart de ces changements.

Souvent, le *locus*[4] où la mutation se produit, se situe dans des zones intergéniques et à ce moment là, la mutation ne produira aucun effet chez l'individu. Cependant, lorsque celle-ci se produit au niveau des gènes, cela crée dans la plupart des cas des différences normales chez les individus porteurs de cette mutation telles qu'une couleur de cheveux ou une couleur des yeux différente. On dit alors que la mutation a créé un nouvel allèle (une version légèrement différente du gène initial). Cependant, il peut arriver que la mutation engendre un nouvel allèle qui compromet la fonction de la protéine qui est codée par ce gène et qui se manifeste par l'apparition de ce que l'on appelle *maladies génétiques.*

Une fois que les nouveaux allèles apparaissent, les recombinaisons génétiques interviennent pour créer davantage de variation génétique.

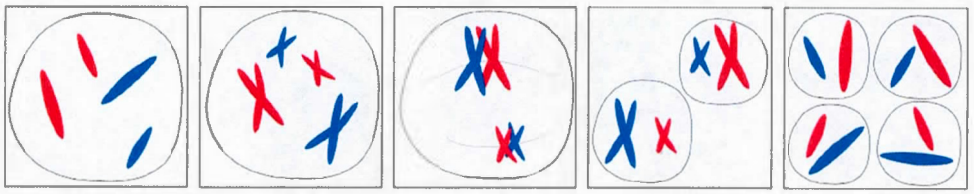
1.2.2 La recombinaison génétique

Le phénomène de recombinaison est très important pour le maintien de la diversité génétique. Il se produit au cours de la méiose un mécanisme de division cellulaire spécifique aux cellules germinales. Ces cellules donneront naissance à de nouvelles cellules appelées gamètes (ce sont les cellules sexuelles : ovules et spermatozoïdes). Les gamètes ont la particularité d'être des cellules haploïdes ou autrement dit elles contiennent seulement la moitié des chromosomes d'une cellule diploïde soit 23 chromosomes. Les différentes étapes de la méiose sont illustrées dans la figure 1.2.

---

3. Source : `http://learn.genetics.utah.edu/content/variation/mutation/`

4. Un locus désigne une position ou un emplacement physique sur un chromosome.



(a) Avant la méiose (b) Réplication (c) Appariement (d) 1^{e} division (e) 2^{e} division

Figure 1.2 Étapes de la méiose : seulement deux paires de chromosomes homologues sont représentés (Tirée de Forest (2010)).

Dans la figure 1.2.a, nous avons représenté deux paires de chromosomes homologues, en bleu ceux hérités du père et en rouge ceux hérités de la mère. Cette figure illustre l'état de la cellule au repos (avant méiose). Les chromosomes sont alors formés d'une seule chromatide[5]. Une fois le processus de division enclenché, chaque chromosome se réplique et devient un chromosome en forme de X constitué de deux chromatides comme l'illustre la figure 1.2.b. Dès que la réplication se termine, il se produit un appariement des chromosomes homologues (figure 1.2.c), et c'est durant cette phase que les deux homologues s'échangent du matériel génétique. C'est ce que l'on appelle la recombinaison intrachromosomique ou cross-over (qui est illustrée sur la figure 1.3). L'endroit où se produit le cross-over (ou enjambement) est aléatoire.

Suite aux enjambements, il se produit une première division de méiose appelée mitose réductionnelle. Grâce à cette première division, nous obtenons deux cellules filles haploïdes qui contiennent chacune 23 chromosomes à deux chromatides, et

5. Une chromatide est une molécule d'ADN associée à des protéines qui a la forme d'un bâtonnet. Les chromatides n'apparaissent sous forme de chromosomes (en forme de X) que durant le processus de division cellulaire. Ainsi, un chromosome en forme de X est composé de deux chromatides.

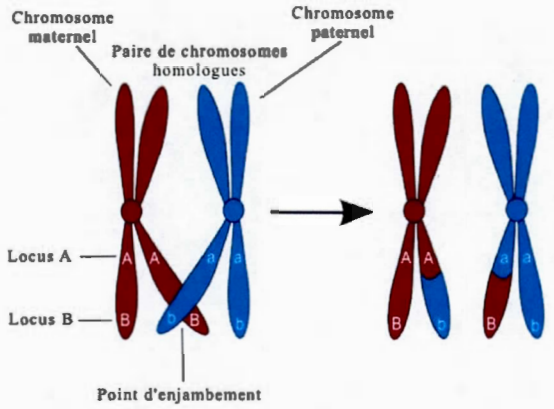


Figure 1.3 Phénomène d'enjambement (cross-over).

c'est également durant cette première division que se fait le brassage entre chromosomes paternels et maternels (recombinaison interchromosomiques), autrement dit, la répartition des chromosomes homologues dans chacune des cellules se fait de façon complètement aléatoire. Par conséquent, sans tenir compte des cross-over, il y a 2^{23} (soit $8,388,608$) gamètes possibles.

Par la suite, une deuxième division de méiose (mitose équationnelle) se produit pour donner naissance à quatre cellules filles contenant chacune 23 chromosomes à une seule chromatide. Tout au long de ce mémoire, lorsque nous parlerons des recombinaisons, en fait, nous faisons référence à la recombinaison intrachromosomique (qui correspond à un nombre pair de cross-over).

1.3 Marqueurs génétiques, haplotype et génotype

Un marqueur génétique est une séquence d'ADN avec un emplacement physique connu sur un chromosome. Plusieurs types de marqueurs existent et les polymorphismes nucléotidiques simples appelés SNP (pour *single nucleotide polymorphism*) sont le type le plus commun de la variation génétique.

Chaque SNP représente une différence d'un seul nucléotide sur la séquence d'ADN.

Par exemple, un SNP peut remplacer le nucléotide cytosine (C) par le nucléotide thymine (T) dans un certain segment d'ADN. En général, les SNPs possèdent juste deux allèles. Alors, une façon de les représenter et que nous allons considérer dans ce mémoire est la notation binaire {0,1}. Ainsi le SNP n'ayant subi aucune mutation au cours du temps aura l'allèle "0", ce dernier est dit *allèle primitif* ou encore *allèle majeur*. Quant au SNP dit *mutant* ou autrement dit, celui qui a subi une mutation aura pour allèle "1" (aussi appelé *l'allèle mineur*).

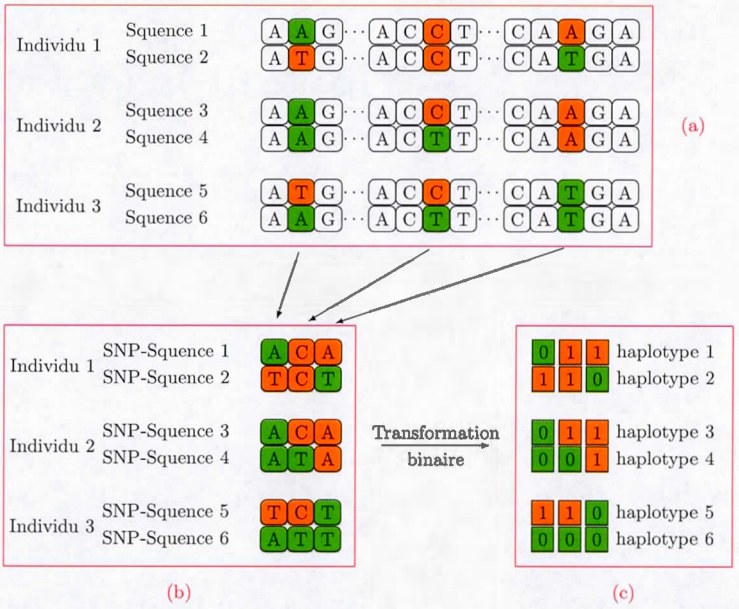


Figure 1.4 Extraction de marqueurs génétiques de type SNP à partir de séquences génétiques et transformation sous forme binaire. Les SNPs mutants sont représentés en orange et les SNPs primitifs en vert.

La figure 1.4.a illustre une partie de six séquences génétiques provenant de trois individus différents. Chaque individu est représenté par deux séquences pour tenir compte de la réalité diploïde. Les SNPs représentés en vert sont regroupés sous forme d'haplotypes (figure 1.4.b) puis transformés sous forme binaire (figure 1.4.c).

De manière générale, on peut définir un haplotype par une suite de loci sur un même chromosome. Ces loci peuvent être des gènes ou des marqueurs génétiques comme ceux illustrés dans la figure 1.4.b.
Le génotype quant à lui, correspond à la composition allélique de chaque locus pour un même individu. Le tableau 1.1 regroupe les génotypes des trois individus aux trois marqueurs de l'exemple de la figure 1.4

Tableau 1.1 Génotype des individus de la figure 1.4.

Individu	Génotype au SNP 1	Génotype au SNP 2	Génotype au SNP 3
1	A|T	C|C	A|T
2	A|A	C|T	A|A
3	T|A	C|T	T|T

1.4 Distance génétique

On peut définir une distance entre deux loci de deux façons. La première est de considérer une distance physique basée sur le nombre de paire de base *pb* (nucléotides) qui les séparent. Cependant, cette mesure peut devenir très contraignante en raison du nombre important de nucléotides qui composent l'ADN ; c'est pourquoi, lorsque l'on considère une région sur un chromosome, on utilise souvent la mégabase (Mb) à la place des paires de base pour décrire la longueur de cette région tel que 1 Mb $= 1,000,000$ *pb*.
L'autre mesure de distance est une mesure génétique basée sur la fraction de recombinaison généralement notée θ ($\theta \leq 1/2$) entre deux loci. Autrement dit, elle représente la probabilité que deux loci sur un même chromosome soient séparés durant la méiose. L'idée est que plus les loci sont proches moins il y aura de chance pour qu'un événement de recombinaison se produise entre eux.

Comme les probabilités sont toujours comprises entre 0 et 1, cela ne permet pas de définir toutes les distances. La solution est alors de définir une distance qui est fonction de la fraction de recombinaison θ. La fonction la plus couramment utilisée est celle définie par Haldane (1919) :

$$d = -\frac{1}{2}\ln(1-2\theta).$$

L'unité de mesure est le *centimorgan* (cM) équivalent à une probabilité d'enjambement de 1% entre deux loci, soit une recombinaison par 100 méioses.
Il est possible de définir une relation entre la distance physique et génétique entre deux loci et qui varie selon l'espèce. Chez l'homme, on admet la relation : 1 cM $\approx$ 1 Mb.

1.5 Équilibre de Hardy-Weinberg et déséquilibre de liaison

1.5.1 Équilibre de Hardy-Weinberg

La notion d'équilibre de Hardy-Weinberg est importante en génétique. Elle permet de montrer sous certaines hypothèses (fortes) que la variabilité génétique sera maintenue au fil des générations. En effet, on pourrait se poser la question si un allèle dominant ne va pas finir par dominer dans toute la population. Pour répondre à cette question, le modèle de Hardy-Weinberg adopte les hypothèses suivantes :

- population panmictique (accouplements aléatoires) de taille infini ;
- générations discrètes ;
- absence de sélection, de migration et de mutation ;
- fréquences génotypiques identiques pour les deux sexes.

Prenons l'exemple d'un marqueur biallélique dont les allèles sont A et a et provenant d'individus d'une population vérifiant les hypothèses ci-haut. Les fréquences correspondant aux trois génotypes possibles AA, Aa et aa dans la population sont respectivement p_{AA}, p_{Aa} et p_{aa} avec $p_{AA}+p_{Aa}+p_{aa}=1$. Ainsi, les fréquences

correspondants aux deux allèles pour la génération actuelle sont alors :

$$\begin{aligned} p_A &= P(A|AA)P(AA) + P(A|Aa)P(Aa) + P(A|aa)P(aa) \\ &= p_{AA} + \frac{1}{2}p_{Aa}. \end{aligned}$$

Par le même raisonnement, on peut montrer que :

$$p_a = p_{aa} + \frac{1}{2}p_{Aa}.$$

Sous l'hypothèse de panmixie, les fréquences génotypiques de la génération suivante sont respectivement :

$$\begin{aligned} p^*_{AA} &= \left(p_{AA} + \frac{1}{2}p_{Aa}\right)^2, \\ p^*_{Aa} &= 2\left(p_{AA} + \frac{1}{2}p_{Aa}\right)\left(p_{aa} + \frac{1}{2}p_{Aa}\right), \\ p^*_{aa} &= \left(p_{aa} + \frac{1}{2}p_{Aa}\right)^2. \end{aligned}$$

Ainsi, en reprenant le même raisonnement pour trouver les fréquences alléliques, on peut montrer aisément que les fréquences à la génération suivante correspondent exactement aux fréquences initiales.

1.5.2 Déséquilibre de liaison

La notion de déséquilibre de liaison (LD pour *linkage disequilibrium*) est importante en cartographie génétique notamment dans les études d'association qu'on discutera un peu plus loin. Notez que la majorité des informations contenues dans cette section proviennent de l'article de Nordborg et Tavaré (2002).

Le LD désigne l'association **non-aléa**toire entre les allèles à des loci différents. Supposons par exemple, que dans une population, nous avons l'allèle a à un certain locus et l'allèle b à un autre locus avec les fréquences p_a et p_b respectivement. Si ces deux loci sont indépendants, alors nous nous attendons à ce que la fréquence de l'haplotype (ab) qu'on note p_{ab} soit égale au produit des deux fréquences p_a et p_b.

Par contre si la fréquence de cet haplotype (ab) dans la population est différente de $(p_a \times p_b)$, alors les deux loci sont dit en LD.

Il existe une variété de mesures statistiques qui permettent d'évaluer le degré d'association entre les paires de marqueurs. Prenons le cas de deux marqueurs biallélіques dont les allèles sont (a, A) et (b, B) situés sur deux loci différents. Une des mesures les plus simples du degré d'association non aléatoire consiste à calculer la différence entre les fréquences observées et espérées d'un haplotype et est défini comme suit :

$$D_{AB} = p_{AB} - (p_A \times p_B).$$

Ainsi, si l'hypothèse que les deux allèles sont indépendants aux deux loci, alors, D_{AB} devrait être significativement différent de zéro.

Une autre mesure très utilisée aussi, est la valeur absolue de D_{AB} normalisée. Elle est donnée par :

$$|D'_{AB}| = \left|\frac{D_{AB}}{D_{\max}}\right| \quad \text{où} \quad D_{\max} = \begin{cases} \min(p_A p_b, p_a p_B) & \text{si} \quad D_{AB} > 0, \\ \min(p_A p_B, p_a p_b) & \text{si} \quad D_{AB} < 0. \end{cases}$$

Une valeur de $|D'_{AB}|$ proche de 1 indique un déséquilibre de liaison tandis qu'une valeur de 0 témoigne d'un équilibre de liaison.

CHAPITRE II

THÉORIE DE LA COALESCENCE

Le processus de coalescence est un processus stochastique qui a été introduit la première fois par Kingman (1982). Ce processus est utilisé en génétique des populations dans le but de retracer la généalogie de manière rétrospective (en arrière dans le temps) ; autrement dit, à partir d'un échantillon de séquences, le processus de coalescence nous permet de remonter à travers les générations jusqu'à trouver l'ancêtre commun le plus récent, communément appelé MRCA (*Most Recent Commun Ancestor*) pour ces séquences.

Ce chapitre a pour but de présenter les notions de base du processus de coalescence qu'on utilisera un peu plus loin (voir chapitre IV) dans la construction de notre modèle. Nous allons donc commencer par décrire le modèle de base de Wright-Fisher qui est basé sur des hypothèses plus ou moins réalistes. Par la suite nous verrons comment affaiblir certaines hypothèses en considérant dans le modèle, des événements tels que les mutations et les recombinaisons génétiques.

2.1 Le modèle de Wright-Fisher

Le modèle introduit par Fisher (1930) et Wright (1931) est un modèle très simple de génétique des populations. Il permet de décrire la relation généalogique entre un ensemble de gènes (ou de séquences). Ce modèle repose sur un certain nombre

d'hypothèses simplificatrices :

1. Générations discrètes qui ne se chevauchent pas : ceci équivaut à supposer que tous les individus de la population ont la même espérance de vie et que la reproduction et la mort se produisent de façon simultanée pour les individus.
2. Pas de sélection naturelle : cela veut dire que tous les individus de la population ont les mêmes chances de se reproduire.
3. Taille de population constante composée d'individus haploïdes.
4. La population n'a pas de structure géographique ou sociale : cela veut dire que les parents associés aux individus d'une génération sont choisis complètement au hasard.
5. Pas de mutation ou de recombinaison génétique.

Pour tenir compte de la réalité diploïde chez l'homme, on considère une population de taille $2N$ séquences haploïdes ($2N$ haplotypes), que l'on pourra voir comme une population de taille N séquences diploïdes (N individus).
Chaque génération issue du modèle de Wright-Fisher est le résultat d'un échantillonnage aléatoire avec remise de la génération précédente. Ainsi, le nombre de descendants pour une séquence i de la génération k est un exemple de loi binomiale de paramètres $n = 2N$ et $p = 1/(2N)$. De ce fait, le nombre moyen de descendants est de 1 descendant par séquence et par génération.
Un exemple de réalisation du processus de Wright-Fisher est illustré à la figure 2.1. Dans cet exemple, nous avons généré une population de taille $2N = 10$ séquences, que nous avons suivi sur 15 générations dans le passé. Chaque ligne horizontale représente une génération constituée de séquences d'ADN, représentées par des petits cercles qui sont liés à leurs ancêtres de la génération précédente. Quatre séquences (représentées en vert (séquences 4, 6, 7 et 10)) ont été échantillonnées à partir de la population des 10 séquences. Il faut remonter de 11 générations pour

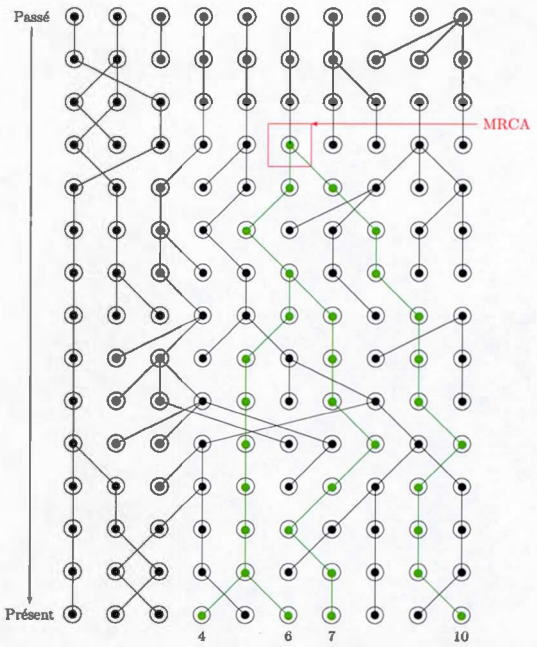


Figure 2.1 Illustration du modèle de Wright-Fisher avec 10 séquences suivies sur 15 générations. Quatre séquences (en vert) ont été échantillonnées. La séquence encadrée en rouge représente le MRCA des séquences échantillonées.

que celles ci trouvent leur ancêtre commun le plus récent.
Il est possible d'isoler la généalogie des quatre séquences échantillonnées et de la représenter sous forme d'arbre tel qu'illustré sur la figure 2.2.
La section suivante traitera notamment de la loi des temps de coalescences qui nous permettra plus tard (voir chapitre IV) de définir une nouvelle mesure de similarité entre individus, nécessaire à la construction de notre modèle.

2.2 Le processus de coalescence

On dit qu'il y a coalescence entre deux séquences de la même génération lorsque celles ci trouvent un ancêtre commun en remontant en arrière dans le temps.

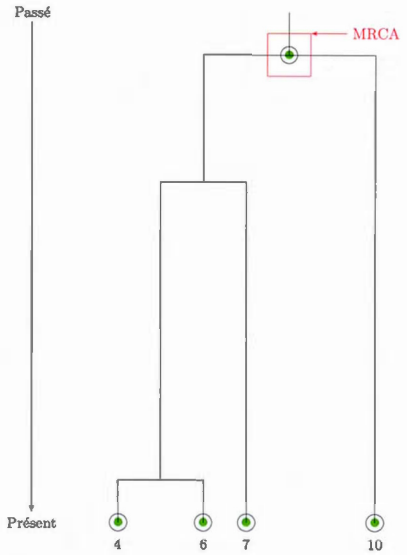


Figure 2.2 Représentation d'un arbre de coalescence obtenu à partir des quatre séquences génétiques échantillonnées de la figure 2.1.

Soit T_2 le temps en nombre de générations avant que deux séquences ne coalescent. La probabilité pour que deux séquences trouvent un ancêtre commun une génération en arrière dans le temps est de $1/(2N)$. En effet, une séquence choisie au hasard un parent de la génération précédente et l'autre séquence aura une chance sur $2N$ de choisir le même parent. De manière générale, le temps avant que deux séquences trouvent un ancêtre commun k générations en arrière est distribué selon une loi géométrique de paramètre $p = 1/(2N)$. Ainsi, nous avons :

$$P(T_2 = k) = \left(1 - \frac{1}{2N}\right)^{k-1} \frac{1}{2N} .$$

Cela veut dire que nos deux séquences trouvent des ancêtres distincts pendant $(k-1)$ générations et coalescent à la $k^{\text{ème}}$ génération. Il faut donc en moyenne remonter $2N$ générations dans le passé pour que deux séquences trouvent un ancêtre commun.

On peut aussi trouver la loi du temps à l'ancêtre commun pour un échantillon composé de n séquences. Pour cela, nous allons tout d'abord calculer la probabilité que n séquences trouvent des ancêtres différents à la génération précédente. Soit $P(n)$ cette probabilité. De ce fait, la première séquence choisie un parent au hasard parmi les $2N$ séquences, par la suite, la seconde séquence choisie un autre parent au hasard parmi les $2N-1$ séquences restantes et ainsi de suite. On aura alors :

$$\begin{aligned}
P(n) &= \left(\frac{2N-1}{2N}\right)\left(\frac{2N-2}{2N}\right)\cdots\left(\frac{2N-n+1}{2N}\right) \\
&= \left(1-\frac{1}{2N}\right)\left(1-\frac{2}{2N}\right)\cdots\left(1-\frac{n-1}{2N}\right) \\
&= \prod_{k=1}^{n-1}\left(1-\frac{k}{2N}\right) \\
&\vdots \\
&= 1-\sum_{k=1}^{n-1}\frac{k}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right) \\
&= 1-\frac{1}{2N}\frac{n(n-1)}{2}+\mathcal{O}\left(\frac{1}{N^2}\right) \\
&= 1-\frac{\binom{n}{2}}{2N}+\mathcal{O}\left(\frac{1}{N^2}\right) \\
&\approx 1-\frac{\binom{n}{2}}{2N},
\end{aligned}$$

où $\mathcal{O}(1/N^2)$ représente tous les termes qui sont divisés par N^i avec $i \in \{2,3,\ldots,n-1\}$. Ainsi, la probabilité d'avoir un événement de coalescence à la génération précédente lorsque l'on a un échantillon composé de n séquences est approximativement égale à $\binom{n}{2}/2N$. Puisque nous avons supposé que les générations se forment de manière indépendante l'une de l'autre, la probabilité qu'aucune coalescence ne surviennent en $(k-1)$ générations, puis qu'une se produise à la $k^{\text{ème}}$ génération est :

$$P(T_n = k) \approx \left(1-\frac{\binom{n}{2}}{2N}\right)^{k-1}\frac{\binom{n}{2}}{2N}.$$

Ainsi, le temps à l'ancêtre commun (T_n) pour un échantillon constitué de n séquences est approximativement distribué selon une loi géométrique de paramètre $p = \binom{n}{2}/2N$.

Il est plus réaliste de considérer le temps de coalescence comme un temps continu, plutôt qu'un temps discret en nombre de générations. Dans le cas où la taille de population est très grande nous pouvons approximer la loi géométrique par une loi exponentielle, où une unité de temps correspondrait au temps moyen pour que deux séquences trouvent un ancêtre commun. Autrement dit, en unité de $2N$ générations.

Proposition *Pour une taille de population assez grande N, le temps à l'ancêtre commun pour un échantillon de n séquences génétiques noté T_n^c est approximativement distribué selon une loi exponentielle de paramètre $\binom{n}{2}$. Ainsi, nous avons :*

$$P(T_n^c > t) \approx \mathrm{e}^{-\binom{n}{2}t}.$$

Preuve *Pour montrer que T_n^c est approximativement distribué selon une loi exponentielle, il suffit de montrer que :*

$$\lim_{N\to+\infty} P(T_n^c > t) = \mathrm{e}^{-\binom{n}{2}t}.$$

Nous savons que le temps à l'ancêtre commun exprimé en nombre de générations est approximativement distribué selon une loi géométrique de paramètre $p = 1/2N$. Nous pouvons donc écrire :

$$\lim_{N\to+\infty} P\left(T_n^c > \frac{k}{2N}\right) = \lim_{N\to+\infty} P(T_n^c > t) = \lim_{N\to+\infty} \left(1 - \frac{\binom{n}{2}}{2N}\right)^{2Nt},$$

où $t = k/2N$. Nous savons aussi que :

$$\lim_{N\to+\infty} \left(1 - \frac{\binom{n}{2}}{2N}\right)^{2Nt} = \lim_{N\to+\infty} \mathrm{e}^{2Nt\ln\left(1-\frac{\binom{n}{2}}{2N}\right)}. \qquad (2.1)$$

Or, nous avons :

$$\begin{aligned} 2Nt \ln\left(1 - \frac{\binom{n}{2}}{2N}\right) &= \frac{\ln\left(1 - \frac{\binom{n}{2}}{2N}\right)}{\frac{1}{2N}} \cdot t \\ &= \frac{\ln\left(1 - \binom{n}{2}x\right)}{x} \cdot t \\ &= \frac{\ln\left(1 - \binom{n}{2}x\right) - \ln\left(1 - \binom{n}{2}(0)\right)}{x - 0} \cdot t, \end{aligned}$$

où $x = 1/2N$. *Ainsi, nous pouvons écrire :*

$$\begin{aligned} \lim_{N \to +\infty} \ln\left(1 - \frac{\binom{n}{2}}{2N}\right)^{2Nt} &= \lim_{x \to 0} \frac{\ln\left(1 - \binom{n}{2}x\right) - \ln\left(1 - \binom{n}{2}(0)\right)}{x - 0} \cdot t \\ &= -\binom{n}{2} t. \end{aligned} \tag{2.2}$$

En remplaçant le résultat de l'équation (2.2) *dans l'équation* (2.1), *nous obtenons finalement :*

$$\lim_{N \to +\infty} P(T_n^c > t) = \mathrm{e}^{-\binom{n}{2}t}.$$

Nous pouvons donc dire que le temps avant le prochain événement de coalescence suit une loi exponentielle de taux $\binom{n}{2}$. Autrement dit, $T_n^c \sim \exp\left(\binom{n}{2}\right)$.

Il est intéressant de trouver le temps avant que toutes nos séquences trouvent un ancêtre commun. Soit T_i^c le temps d'attente pour que l'on ait une coalescence alors que l'on a i séquences dans la généalogie. Le temps total jusqu'à ce que l'on trouve l'ancêtre commun le plus récent (MRCA) est donné par :

$$T_{MRCA} = \sum_{i=2}^{n} T_i^c.$$

La loi de T_{MRCA} peut être obtenue par une convolution de variables exponentielles (Hein *et al.*, 2004). Cependant, l'espérance de T_{MRCA} peut être facilement obtenue

comme suit :

$$\begin{aligned}
\mathbb{E}(T_{MRCA}) &= \mathbb{E}\left(\sum_{i=2}^{n} T_i^c\right) \\
&= \sum_{i=2}^{n} \mathbb{E}(T_i^c) \\
&= \sum_{i=2}^{n} \frac{1}{\binom{i}{2}} \\
&= \sum_{i=2}^{n} \frac{2}{i(i-1)} \\
&= 2\sum_{i=2}^{n} \left(\frac{1}{i-1} - \frac{1}{i}\right) \\
&\vdots \\
&= 2\left(1 - \frac{1}{n}\right).
\end{aligned}$$

On constate que le temps total moyen pour remonter à l'ancêtre commun le plus récent de toutes les séquences est inférieur à 2 en unité de $2N$ générations, et que la coalescence de la dernière paire de séquences prend en moyenne la moitié du temps total. Ce résultat ne semble pas très intuitif lorsqu'on le voit pour la première fois. En effet, lorsqu'on augmente la taille de l'échantillon cela n'apporte guère plus d'information et ne fait que rajouter des petites branches en bas de l'arbre, parce-qu'en réalité il y a une seule histoire qui s'est produit au cours du temps.

À présent nous sommes en mesure de décrire un premier algorithme qui permet de simuler un processus de coalescence pour un échantillon de n séquences.

Algorithme 1 :

1. Poser $i = n$ séquences.
2. Simuler le temps T_i^c avant le prochain événement, $T_i^c \sim \exp\left(\binom{i}{2}\right)$.
3. Choisir une paire de séquence de façon uniforme parmi les $\binom{i}{2}$ paires possibles.

4. Fusionner les deux séquences choisies en une seule séquence et diminuer la taille de l'échantillon de un. Autrement dit, $i \leftarrow i - 1$.

5. Si $i > 1$, aller en 2.

2.3 Le processus de coalescence avec mutation

Nous avons vu dans la section 1.2.1 du chapitre I que les mutations étaient en partie responsables de diversité génétique. Ainsi, pour avoir un modèle plus proche de la réalité, nous devons prendre en considération la possibilité qu'il puisse y avoir des mutations qui se produisent dans la généalogie des séquences. C'est cela que cette section traitera essentiellement de l'ajout des événements de mutation dans le processus de coalescence.

Il existe plusieurs modèles avec lesquels les événements de mutations peuvent être considérés, notamment le modèle de "sites infinis" et le modèle "d'allèles infinis". Le modèle d'allèles infinis suppose que l'information disponible sur les allèles permet seulement de dire si ils sont identiques ou pas. Une mutation va donc créer un nouvel allèle non observé auparavant.
Quant au modèle de sites infinis (que nous allons considérer tout au long de ce mémoire), il suppose qu'une mutation ne se produit qu'une seule fois au même endroit sur la séquence d'ADN au cours de l'histoire. Autrement dit, la probabilité qu'une mutation apparaisse plusieurs fois sur le même locus (site) est négligeable. Il est à noter que lorsqu'un événement de mutation se produit sur une séquence, tous les descendants du porteur de cette mutation la posséderont aussi.

En considérant un modèle neutre où la mutation qui se produit le long d'une séquence n'a aucun effet sur les probabilités de décès ou de reproduction, alors le processus de mutation est indépendant du processus de coalescence. De ce fait, les mutations peuvent être superposées directement sur l'arbre de coalescence. La

figure 2.3 illustre un arbre issu de la réalisation d'un processus de coalescence avec mutation.

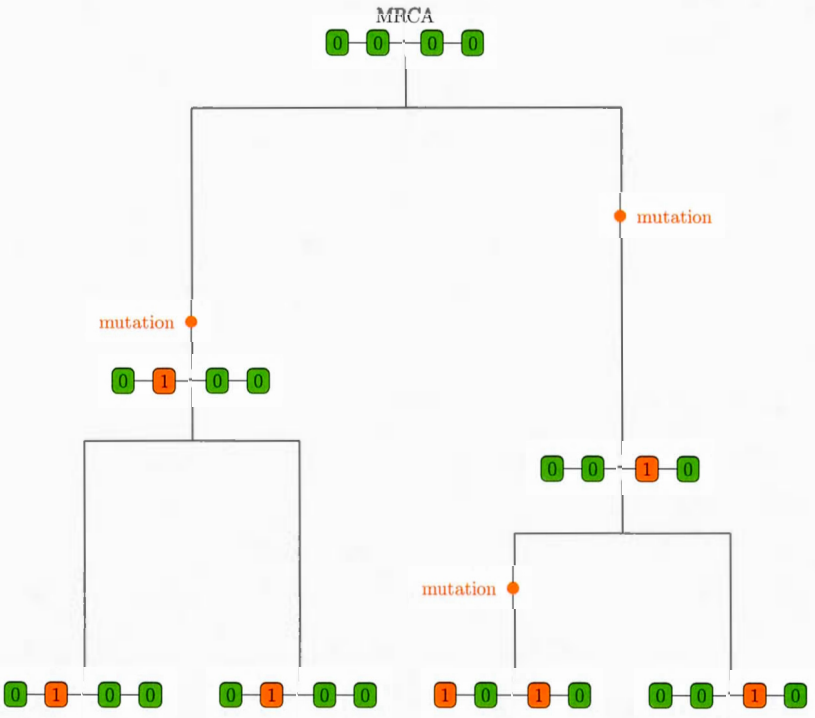


Figure 2.3 Représentation d'un arbre de coalescence avec mutation obtenu à partir de quatre séquences génétiques composées de quatre marqueurs. Un carré vert représente un allèle primitif tandis qu'un carré orange représente un allèle mutant (ou dérivé).

À présent, intéressons nous au modèle mathématique. Soit μ la probabilité d'avoir une mutation à un locus d'une séquence pour une génération donnée. Le temps T^M en nombre de générations avant qu'un événement de mutation ne se produise est distribué selon une loi géométrique de paramètre μ. Ainsi, nous pouvons écrire :

$$P(T^M = k) = (1 - \mu)^{k-1}\mu.$$

De façon analogue que dans le processus de coalescence sans mutation, on peut approximer la distribution du temps T^M par la loi exponentielle lorsque N est assez grand. Ainsi, nous avons :

$$P(T^M > t) = (1-\mu)^{2Nt} = \left(1 - \frac{\theta/2}{2N}\right)^{2Nt} \approx \mathrm{e}^{-\frac{\theta}{2}t},$$

où le paramètre $\theta = 4N\mu$ représente le taux de mutation de la population.
On considérant n lignées indépendantes, le temps avant d'observer une mutation sur une lignée est de loi exponentielle de paramètre $n\theta/2$. Ainsi, le temps d'attente avant le prochain événement (coalescence ou mutation) $T = \min\{T_n^c, T_n^M\}$ est distribué selon une loi exponentielle de paramètre $\binom{n}{2} + n\theta/2$. En effet,

$$\begin{aligned} P(T \leq t) = P\{\min(T_n^c, T_n^M) \leq t\} &= 1 - P\{\min(T_n^c, T_n^M) > t\} \\ &= 1 - P(T_n^c > t, T_n^M > t) \\ &= 1 - \left[P(T_n^c > t) \cdot P(T_n^M > t)\right] \\ &= 1 - e^{-t\binom{n}{2}} e^{-t\frac{n\theta}{2}} \\ &= 1 - \mathrm{e}^{-t\left(\binom{n}{2} + \frac{n\theta}{2}\right)}. \end{aligned}$$

La probabilté que le prochain événement soit une coalescence correspond à $P(T_n^c \leq T_n^M)$. Cette probabilité peut être obtenue comme suit :

$$\begin{aligned} P(T_n^c \leq T_n^M) &= \int_0^\infty \int_0^{t_2} f_{T_n^c, T_n^M}(t_1, t_2)\, \mathrm{d}t_1 \mathrm{d}t_2 \\ &= \int_0^\infty \int_0^{t_2} f_{T_n^c}(t_1) f_{T_n^M}(t_2) \mathrm{d}t_1 \mathrm{d}t_2 \qquad (T_n^c \perp\!\!\!\perp T_n^M) \\ &= \int_0^\infty \int_0^{t_2} \binom{n}{2} \mathrm{e}^{-t_1\binom{n}{2}} \frac{n\theta}{2} \mathrm{e}^{-\frac{n\theta}{2}t_2} \mathrm{d}t_1 \mathrm{d}t_2 \\ &\vdots \\ &= \frac{\binom{n}{2}}{\binom{n}{2} + \frac{n\theta}{2}} \\ &= \frac{n-1}{n+\theta-1}. \end{aligned}$$

De façon analogue, la probabilité que le prochain évènement soit une mutation est donc égale à $\theta/(n+\theta-1)$.

2.4 Le processus de coalescence avec recombinaison

Dans cette section, nous allons nous concentrer sur l'ajout d'un événement de recombinaison dans le processus de coalescence.
C'est à Hudson (1983) qu'on doit l'introduction de la recombinaison dans le processus de coalescence. Hudson a présenté un modèle très simple de processus de coalescence avec recombinaison. Le principe est illustré sur la figure 2.4.

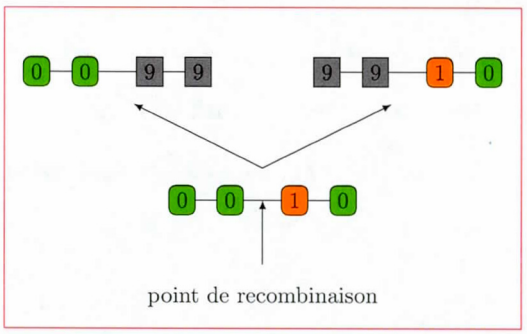


Figure 2.4 La recombinaison selon le modèle d'Hudson. À partir d'une séquence composée de quatre marqueurs ancestraux représentés par les carrés verts et oranges, nous obtenons deux séquences parentales composées de marqueurs ancestraux ainsi que de marqueurs non-ancestraux représentés par des carrés gris.

Lorsque l'on décrit les événements de recombinaison du passé vers le présent, c'est à partir de deux séquences parentales que nous obtenons une nouvelle séquence constituée d'un mélange des deux premières (séquences parentales) à gauche et à droite du point de recombinaison. Cependant, puisque le processus de coalescence construit la généalogie du présent vers le passé, il nous est impossible de connaître exactement la composition de toute la séquence parentale mais uniquement d'une partie. Toutefois, cela ne cause aucun problème : en effet, les séquences parentales vont être complétées par de la matière "non-ancestrale" situé à gauche ou à droite

du point de recombinaison. La présence de matière non ancestrale ne cause aucun problème vu que cette partie ne se trouve pas dans les séquences de notre échantillon, ce qui la rend non informative pour notre généalogie.
Il est à noter que la réalisation du processus avec recombinaison ne donne plus un arbre mais un graphique complexe appelé "graphe de recombinaison ancestral" ou ARG pour "*Ancestral Recombination Graph*", tel qu'illustré sur la figure 2.5. Même si ce modèle ne rend pas tout à fait compte de toute la complexité biologique de la recombinaison présenté à la section 1.2.2, il forme toujours la base pour la plupart des applications en théorie de la coalescence.

A présent, nous allons présenter le modèle mathématique qui permet de modéliser la recombinaison. Soit r la probabilité d'avoir une recombinaison le long d'une séquence à une génération donnée. Si l'on considère un modèle discret alors le temps en nombre de générations jusqu'à ce qu'un événement de recombinaison se produise est distribué selon une loi géométrique de paramètre r. Ainsi, la probabilité d'avoir une recombinaison k générations en arrière dans le temps est :

$$P(T^R = k) = (1-r)^{k-1}r.$$

De façon similaire au cas du processus de coalescence avec mutation, on définit $\rho = 4Nr$, le taux de recombinaison de la population. On aura alors $r = (\rho/2)/2N$ et l'on peut alors faire une approximation en temps continu de la même façon décrite à la section précédente et montrer que : $T^R \sim \exp(\rho/2)$.
En considérant un échantillon de n séquences et des lignées indépendantes les unes des autres, le prochain événement de recombinaison est alors distribué selon une loi exponentielle de paramètre $(n\rho/2)$. Autrement dit : $T_n^R \sim \exp(n\rho/2)$.
Lorsque l'on introduit les événements de recombinaison, on pourrait se demander si l'on parvient toujours à trouver un ancêtre commun pour l'ensemble de notre échantillon vu qu'à chaque événement de recombinaison le nombre de séquence augmente de 1. La réponse est oui : en effet, le taux "de coalescence" est un

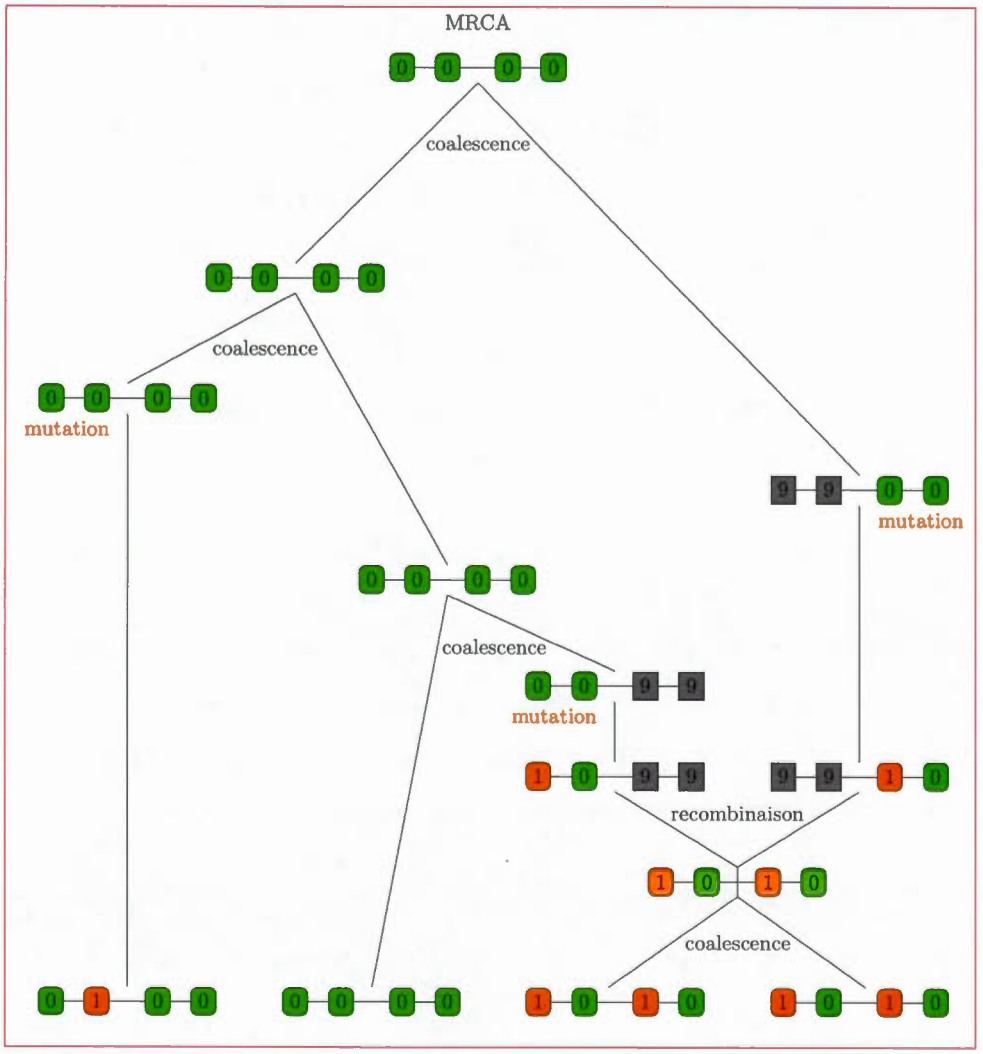


Figure 2.5 Graphe de recombinaison ancestral obtenu à partir de quatre séquences génétiques composées de quatre marqueurs. Un carré vert représente un allèle normal tandis qu'un allèle mutant est représenté par un carré orange. Les carrés gris représentent des marqueurs non-ancestraux dus aux événements de recombinaisons.

taux quadratique en n (le nombre de séquences), et est donc supérieur au taux de recombinaison qui est "linéaire" en n.

Puisque que nous avons une indépendance entre les événements de coalescence, de mutation et de recombinaison, on peut montrer que la distribution du temps avant le prochain événement suit une loi exponentielle de paramètre $n(n+\theta+\rho-1)/2$. Les probabilités associées aux trois événements sont $(n-1)/(n+\theta+\rho-1)$, $\theta/(n+\theta+\rho-1)$ et $\rho/(n+\theta+\rho-1)$ pour une coalescence, une mutation et une recombinaison respectivement.

L'algorithme suivant résume de façon générale comment simuler un ARG.

Algorithme 2 :

1. Poser $i = n$ séquences.
2. Simuler le temps T_i avant le prochain événement, $T_i \sim \exp\left(\frac{i(i+\theta+\rho-1)}{2}\right)$.
3. Le prochain événement sera respectivement un événement de coalescence, de mutation ou de recombinaison avec les probabilités : $(i-1)/(i+\theta+\rho-1)$, $\theta/(i+\theta+\rho-1)$ et $\rho/(i+\theta+\rho-1)$.
4. — Si c'est un événement de coalescence, choisir deux séquences au hasard de façon uniforme parmis les séquences qui peuvent coalescer et les fusionner en une seule puis diminuer la taille de l'échantillon de un. Autrement dit : $i \leftarrow i-1$.
   — Si c'est un événement de mutation, choisir aléatoirement une lignée et un locus pour apposer une mutation.
   — Si c'est un événement de recombinaison, choisir aléatoirement une séquence et un point de recombinaison de manière uniforme le long de la séquence puis créer deux nouvelles séquences telles qu'illustré à la figure 2.4 et augmenter le nombre de séquence de un. Autrement dit : $i \leftarrow i+1$.
5. Si $i > 1$, aller en 2.

CHAPITRE III

CARTOGRAPHIE GÉNÉTIQUE

Le chapitre II nous a permis d'introduire les bases théoriques du processus de coalescence qui sert à la construction de notre nouvelle mesure de similarité entre les individus de notre échantillon, et dont nous présenterons tous les détails dans le chapitre IV. Dans ce troisième chapitre, nous allons plutôt nous intéresser à quelques modèles utilisés en cartographie génétique, et dont nous nous sommes inspirés pour créer notre modèle.

On peut dire que la cartographie des gènes étudie la relation entre les génotypes et les phénotypes. Son objectif est d'identifier le plus précisément possible quelles régions génomiques peuvent affecter un phénotype d'intérêt, mais aussi d'estimer l'importance de ces régions dans la variabilité phénotypique du trait.

Les phénotypes peuvent correspondre au statut de la maladie pour des plans d'études en cas-témoins (généralement codés 0 ou 1), ou à des mesures quantitatives associées à un individu, telles que la pression artérielle ou le taux de glycémie à jeun par exemple.

On peut distinguer deux types d'analyses qui permettent d'estimer la position d'un gène responsable de l'apparition d'un phénotype d'intérêt chez un individu. Ces analyses diffèrent l'une de l'autre, notamment par le type d'échantillon considéré dans l'étude.

Les analyses de liaison (ou *linkage analysis* en anglais) sont en général effectuées sur des échantillons de familles dont l'un ou plusieurs membres sont atteints d'une certaine maladie. L'idée est alors de tester la co-ségrégation entre la maladie et les allèles de marqueurs à proximité, qui ont tendance à être hérités ensemble lors de la méiose entre les membres affectés et non affectés d'une même famille. Cette méthode va en général conduire à la localisation d'une région chromosomique ou, plus rarement, d'un seul gène. L'avantage principal de travailler sur des échantillons de familles est qu'en général les membres d'une même famille partagent des environnements et des gènes similaires. Cependant, le nombre de familles recrutées pour ce genre d'étude est généralement trop faible pour conclure efficacement.
Les analyses d'association de leur coté sont généralement réalisées sur la base d'échantillons provenant d'une certaine population dont les individus sont supposés n'avoir aucun lien de parenté entre eux. Cette fois, l'idée est de tester l'association (ou la corrélation) entre un phénotype d'intérêt avec le génotype au marqueur de l'individu. On peut distinguer entre deux types d'approches pour réaliser de telles analyses. L'approche gène-candidat teste la présence d'une association entre un phénotype d'intérêt et un certain gène dont on soupçonne l'implication dans l'expression du phénotype. Ainsi, cette approche repose sur une bonne connaissance de la fonction du gène mis en examen. La seconde approche consiste quant à elle à la recherche d'association sur l'ensemble du génome. Ces études sont connues sous le nom de *Genome Wide Association Studies (GWAS)* qui sont rendues possible, notamment grâce au développement de nouvelles technologies de séquençage et de génotypage.

Les sections qui suivent vont développer plus en détails les méthodes statistiques qui permettent de réaliser de telles études. Nous allons donc faire un très bref survol sur les études de liaison et nous nous concentrerons particulièrement sur les tests utilisés dans les études d'association. Mais avant, nous donnerons un

bref aperçu sur les maladies complexes ainsi que l'hypothèse sous-jacente à la réalisation de *GWAS*.

3.1 Maladies complexes

À la différence des maladies Mendéliennes classiques qui résultent d'une mutation dans un seul gène, les maladies dites complexes ou multifactorielles (comme la maladie de l'asthme, de Parkinson et du diabète) sont, quant à elles, causées par l'interaction de multiples facteurs. Ces derniers peuvent être génétiques, environnementaux ou même correspondre à l'hygiène de vie de la personne.
La difficulté avec ce genre de maladie est que même si une personne est prédisposée génétiquement à avoir de telles maladies, cela ne veut pas forcément dire que cette personne développera effectivement la maladie au cours de sa vie. Ainsi, plutôt que d'étudier les facteurs génétiques et environnementaux séparément, les chercheurs étudient maintenant comment ces derniers interagissent les uns avec les autres dans le but d'améliorer la compréhension des éventuelles causes de ces maladies. En ce qui concerne le facteur de risque génétique lié à ces maladies, deux grandes hypothèses ont été proposées par les chercheurs. L'hypothèse de maladie commune-variants communs (*common disease-common variant*, CDCV) stipule que la maladie est, entre autres, causée par l'accumulation des effets de plusieurs variants génétiques communs et dont la fréquence allélique dans la population est supérieure ou égale à 5%, et qui ont chacun une petite contribution dans l'apparition de la maladie (Schork *et al.*, 2009). À l'opposé, l'hypothèse de maladie commune-variants rares (*common disease-rare variant*, CDRV) suppose que les variants génétiques rares dont la fréquence allélique est inférieure à 5% dans la population ont une grande influence sur la maladie (Schork *et al.*, 2009). Plusieurs tests d'association considérant l'une ou l'autre de ces hypothèses ont été développés dans la littérature, et nous en présenterons quelques uns dans les sections 3.3 et 3.5.

3.2 Analyse de liaison

Tel qu'il a été mentionné précédemment, l'analyse de liaison est une étude statistique basée sur des données d'individus apparentés (une famille). Chaque famille peut être représentée sous la forme d'un *pedigree*. Ce dernier représente un digramme décrivant les liens de parenté entre les individus d'une même famille un peu comme un arbre généalogique tel qu'illustré sur la figure 3.1.

On se souvient que le paramètre qui mesure la liaison entre deux loci est la frac-

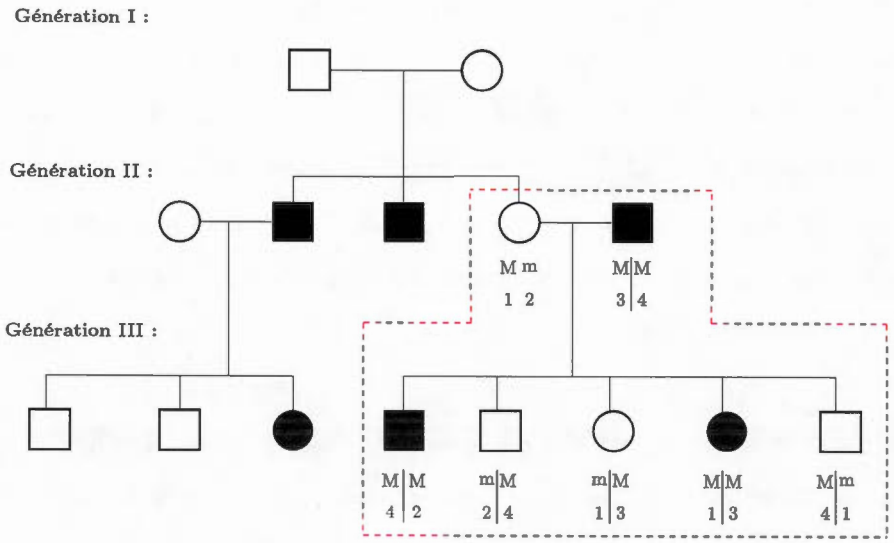


Figure 3.1 Représentation d'un pedigree constitué de trois générations. Les individus de sexe féminin sont représentés par des cercles tandis que ceux de sexe masculin par des carrés. Un cercle ou un carré vide indique que l'individu n'est pas atteint par la maladie, tandis qu'un symbole remplis indique que l'individu est atteint par la maladie.

tion de recombinaison θ décrite au chapitre I. On se rappelle aussi que si deux loci sont liés (ou proche l'un de l'autre), alors la probabilité qu'il y ait une recom-

binaison entre eux est plus petite que 0.5. Ainsi, dans une analyse de liaison on cherche à tester l'hypothèse d'absence de liaison contre l'hypothèse de présence de liaison qui se traduit formellement par :

$$\mathrm{H}_0 : \theta = 0.5 \quad \text{contre} \quad \mathrm{H}_a : \theta < 0.5.$$

En général, le véritable mode de transmission de la maladie est inconnu. C'est pourquoi, avant d'effectuer une analyse de liaison, il est nécessaire de spécifier au préalable un modèle génétique de la maladie. Pour donner une idée sur les différents modèles génétiques qui existent, considérons une maladie monogénique (où un seul gène est impliqué) avec deux allèle possibles M (allèle mutant ou causal) et m (allèle normal).

On définit le vecteur de pénétrances par $F = (f_0, f_1, f_2)$, où f_i correspond à la probabilité d'être malade conditionnellement au génotype de l'individu. Autrement dit,

$$f_0 = P(\text{Malade} \mid mm), \quad f_1 = P(\text{Malade} \mid Mm), \quad f_2 = P(\text{Malade} \mid MM).$$

Un modèle où la maladie n'apparaît que si l'individu possède deux allèles mutants M est dit récessif. Ce dernier conduit aux pénétrances $f_0 = f_1 = 0$. À l'opposé, un modèle dominant nécessite la présence d'un seul allèle mutant et conduit aux pénétrances $f_1 = f_2 = 1$. Lorsque $f_0 > 0$, il y a phénocopie, ce qui veut dire que les facteurs qui expliquent cette maladie ne sont pas seulement génétiques, mais qu'il pourrait y avoir un facteur de risque, environnemental par exemple, qui entraîne l'apparition de la maladie ; c'est, entre autres, ce qui arrive avec les maladies complexes (ou multifactorielles) décrites précédemment.

Afin de tester l'hypothèse de liaison citée plus haut, nous allons introduire quelques notations. Dans un cas simple où l'on pourrait compter le nombre r de recombinants et le nombre $n-r$ de non recombinants lorsque l'on a n méioses informatives (c'est à dire, une méiose où l'on pourrait savoir s'il y eu recombinaison ou non),

la fonction de vraisemblance correspond simplement à la fonction de masse de la loi binomiale de paramètres n et θ et est donnée par :

$$L(\theta; R) = \binom{n}{r} \theta^r (1-\theta)^{n-r}.$$

La valeur de θ qui maximise la vraisemblance est $\theta = r/n$. Cependant, puisque $\theta \in [0, 1/2]$, l'estimateur du maximum de vraisemblance de θ est donné par :

$$\hat{\theta} = \begin{cases} \frac{r}{n} & \text{si} \quad r \leq n/2, \\ \frac{1}{2} & \text{si} \quad r > n/2. \end{cases}$$

Une fois θ estimé, il possible de tester l'hypothèse liaison en utilisant le LOD score. Ce dernier a été introduit par Haldane et Smith (1947) et se définit par :

$$LOD(\theta) = Z(\theta) = \log_{10} \left(\frac{L(\theta; R)}{L(\theta = 0.5; R)} \right),$$

où $L(\theta; R)$ désigne la fonction de vraisemblance calculée à partir du pedigree. Ainsi, le maximum du LOD score sera $Z(\hat{\theta})$. Lorsque $Z(\hat{\theta}) > 3$, cela indique une présence de liaison, si par contre, $Z(\hat{\theta}) < -2$, cela voudrait dire que les deux loci ne sont pas liés. Enfin, si $-2 < Z(\hat{\theta}) < 3$, alors on ne peut pas conclure sur la présence ou non de liaison et des analyses supplémentaires doivent être réalisées (Teare et Barrett, 2005).

Pour illustrer le calcul du LOD score, prenons l'exemple simple d'une partie du pedigree (encadré en rouge) de la figure 3.1. On suppose que la maladie est récessive avec $F = (0, 0, 1)$ et l'allèle causal noté par M. Le père est homozygote au locus de la maladie, il n'est donc pas informatif pour la méiose. La mère est hétérozygote au deux loci, nous avons donc 5 méioses informatives. Cependant, les phases du génotype ne sont pas connues (on ne sait pas quel allèle se trouve sur quel haplotype). Nous allons alors considérer les deux phases possibles pour la mère : $Ph_1 = (M1 \mid m2)$ et $Ph_2 = (M2 \mid m1)$ avec $P(Ph_i) = 1/2$ pour tout

$i \in \{1, 2\}$. La vraisemblance s'écrit alors comme suit :

$$L(\theta) = \sum_{i=1}^{2} P(Ph_i) \cdot L(\theta \mid Ph_i).$$

Si Ph_1 est valide, on compte 3 recombinants et 2 non recombinants. Si Ph_2 est valide, on compte 2 recombinants et 3 non recombinants. On aura alors :

$$\begin{aligned} L(\theta) &= \sum_{i=1}^{2} P(Ph_i) \cdot L(\theta \mid Ph_i) \\ &= \frac{1}{2}\left[\binom{5}{3}\theta^3(1-\theta)^2 + \binom{5}{2}\theta^2(1-\theta)^3\right] = 5\theta^2(1-\theta)^2. \end{aligned}$$

Sur l'intervalle $[0, 1/2]$, la vraisemblance atteint son maximum pour $\hat{\theta} = 1/2$. De ce fait :

$$Z(\hat{\theta}) = \log_{10} \frac{L(1/2)}{L(\theta = 1/2)} = \log_{10}(1) = 0.$$

Ainsi, dans ce cas, on ne peut conclure une liaison, et des études supplémentaires avec d'autres familles doivent être faites.

L'analyse que nous venons de décrire est une *analyse deux points* ou, autrement dit, elle utilise seulement un marqueur de position connu pour tester la liaison. Il est aussi possible de considérer plusieurs marqueurs au lieu d'un seul. Cette analyse est appelée *analyse multipoints*, qui repose également sur le même principe que nous venons de décrire pour une *analyse deux points.*

D'autres approches non-paramétriques que nous ne discuterons pas ici peuvent être employées pour une analyse de liaison ; le lecteur intéressé pourra se référer à Holmans (2001) pour plus de détails.

3.3 Études d'association

De façon générale, les études d'association cherchent à détecter une association entre un variant génétique et la maladie au niveau populationnel. Ainsi, pour les

études en cas-témoins par exemple, elles permettent de nous informer qu'un allèle spécifique d'un gène se retrouve plus fréquemment qu'attendu chez un groupe d'individus affectés (cas), en comparaison à un groupe d'individus non-affectés (témoins ou contrôles). Deux types d'association peuvent avoir lieu comme l'illustre la figure 3.2. On dit qu'il y a une association directe si le marqueur observé est

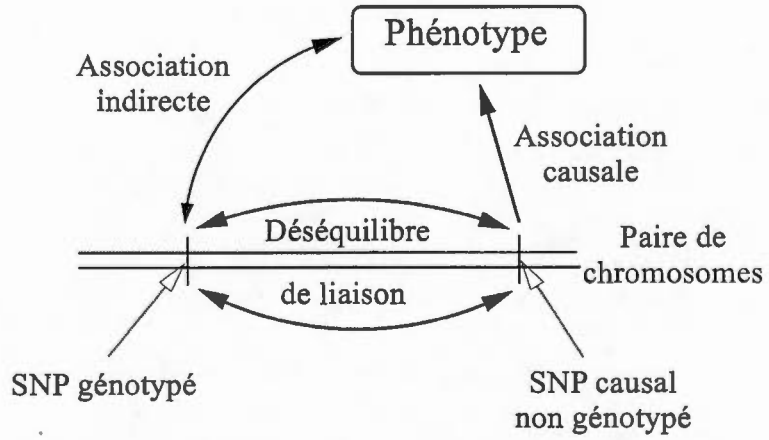


Figure 3.2 Type d'association entre le marqueur et le phénotype (Adaptée de (Astle et Balding, 2009)).

un locus responsable de l'apparition du phénotype (la maladie), ou indirecte si ce marqueur se trouve physiquement proche du locus responsable de la maladie et que leurs allèles sont statistiquement associés en raison du déséquilibre de liaison (LD). On se souvient que le LD reflète l'association non aléatoire, entre les allèles de deux ou plusieurs loci génétiques.

Nous allons à présent décrire les principaux tests statistiques cités dans la littérature, à savoir les tests basés sur les tableaux de contingence ainsi que les tests basés sur les modèles linéaires.

3.3.1 Tests basés sur les tableaux de contingence

Les tests d'association génétique basés sur les tableaux de contingence sont habituellement effectués pour chaque SNP. Ainsi, pour chaque marqueur biallélique avec l'allèle mineur M (ou mutant) et l'allèle majeur m (ou normal), un tableau de contingence peut être construit (voir tableau 3.1). Ce tableau contient le nombre de cas et de contrôles observés pour chaque génotype (mm, Mm et MM).

Sous l'hypothèse nulle de non association avec la maladie, on s'attend à ce que les

Tableau 3.1 Tableau de contingence génotypique.

	Génotype mm	Génotype Mm	Génotype MM	$\sum$
Cas	n_{11}	n_{12}	n_{13}	n_{cas}
Contrôles	n_{21}	n_{22}	n_{23}	$n_{\text{contrôle}}$
$\sum$	n_{mm}	n_{Mm}	n_{MM}	n

fréquences génotypiques soient identiques chez les malades et les non-malades. De ce fait, un test d'association peut être réalisé par un simple test d'indépendance du khi-deux (χ^2) afin de tester l'indépendance entre les lignes et les colonnes du tableau. La statistique de test utilisée est

$$\chi^2 = \sum_{i=1}^{6} \frac{(O_i - E_i)^2}{E_i} \overset{H_0}{\sim} \chi^2(2)$$

qui, sous H_0, suit une loi de khi-deux (χ^2) à deux degrés de liberté[1] (ddl). Les O_i représente les valeurs observées $n_{11}, n_{12}, \ldots, n_{23}$ et les E_i les valeurs attendues sous l'hypothèse nulle (par exemple $E_1 = n_{aa} \cdot n_{cas}/n$) .

Le test classique du χ^2 est avantageux computationnellement mais n'est qu'asymptotique. C'est pourquoi, lorsque les effectifs sont faibles ($n_{ij} < 5$), on utilise plutôt

1. Le degré de liberté (ddl) correspond à $(n-1) \times (d-1)$ où n est le nombre de lignes du tableau et d le nombre de colonnes.

le test exact de Fisher. Ce dernier permet de calculer la probabilité exacte d'observer le tableau 3.1 sous l'hypothèse nulle d'indépendance à l'aide de la distribution hypergéométrique au lieu d'utiliser une approximation comme le test du khi-deux. Ces tests peuvent aisément être réalisés avec le logiciel R par exemple.
Il est à noter que le tableau 3.1 n'est qu'un exemple de la façon dont on pourrait construire un tableau de contingence pour une analyse d'association. En effet, on peut construire un tableau de différentes façons en fonction du modèle génétique (multiplicatif[2], additif, récessif ou dominant) que l'on veut tester.
Lorsque l'on considère un modèle multiplicatif, par exemple, il est nécessaire de construire des tableaux alléliques (2×2) où l'on dénombre les cas et les contrôles pour chaque allèle plutôt qu'une table génotypique (Clarke *et al.*, 2011).
Ainsi, un test d'association allélique peut être réalisé avec un test de χ^2 à 1 ddl ou un test exact de Fisher. Pour tester l'association en supposant un modèle dominant (ou récessif respectivement), il suffit de regrouper les génotypes Mm et MM (respectivement mm et Mm) dans une même cellule du tableau 3.1.

3.3.2 Modèles linéaires

À la différence des tests basés sur les tableaux de contingence décrits à la section précédente, les modèles linéaires permettent non seulement de tester l'effet du génotype sur un phénotype binaire ou continu mais aussi de fournir une estimation de la force d'association entre le trait et le génotype en plus de la possibilité d'inclure des covariabes tel que l'âge ou le sexe de l'individu. Pour l'instant, supposons un modèle simple sans covariables où l'on considère uniquement l'influence d'un seul marqueur génétique situé au locus l sur le phénotype $\boldsymbol{Y}$. Si ce dernier est

---

2. Un modèle multiplicatif indique que le risque de développer la maladie est multiplié par γ pour chaque allèle M additionnel (où γ est un paramètre de pénétrance, $\gamma > 1$), voir Clarke *et al.* (2011) pour plus de détails.

continu, le modèle s'écrit sous la forme suivante :

$$Y_i = \mu_0 + \beta_l G_{il} + \epsilon_i, \tag{3.1}$$

où μ_0 est une constante et $G_{il} \in \{0, 1, 2\}$ est le génotype de l'individu i au marqueur situé au locus l. Le génotype est codé de façon à compter le nombre d'allèles mineurs que l'individu possède au locus l. Ainsi, le coefficient β_l représente l'effet du génotype ($\boldsymbol{G_l}$) sur le phénotype $\boldsymbol{Y}$. Enfin, ϵ_i est un terme d'erreur de moyenne nulle et de variance σ^2_ϵ. Lorsque le phénotype est binaire comme dans les études cas-témoins, on considère plutôt un modèle de régression logistique :

$$\text{logit}\{P(Y_i = 1 \mid G_{il})\} = \mu_0 + \beta_l G_{il}. \tag{3.2}$$

Une fois le modèle de l'équation (3.1) ou (3.2) ajusté, on peut tester l'hypothèse nulle de non-association contre l'hypothèse d'association qui se traduit formellement par :

$$\mathrm{H}_0 : \beta_l = 0 \quad \text{contre} \quad \mathrm{H}_a : \beta_l \neq 0.$$

Le test peut être effectué en utilisant la statistique de Wald donnée par

$$W = \frac{\hat{\beta}_l^{\,2}}{\text{Var}(\hat{\beta}_l)} \overset{\mathrm{H}_0}{\sim} \chi^2_1,$$

où $\hat{\beta}_l$ est l'estimateur du maximum de vraisemblance de β_l. Sous l'hypothèse nulle, W suit approximativement une loi de khi-deux à un degré de liberté.

Le modèle linéaire simple utilisant un seul SNP à la fois a permis l'identification de loci impliqués dans des maladies monogéniques (où un seul gène est impliqué), mais s'est révélé inefficace pour les maladies complexes où l'on soupçonne l'implication de multiples loci. Une autre approche consiste à analyser simultanément un ensemble de marqueurs génétiques de façon à évaluer l'influence de chaque SNP sur le phénotype en présence de tous les autres marqueurs. Ainsi, pour tester l'association entre le phénotype $\boldsymbol{Y}$ et une région $\boldsymbol{G_l} = (\boldsymbol{G_{1l}}, \ldots, \boldsymbol{G_{pl}})$ constituée de p

de marqueurs génétiques située au locus l, on commence par ajuster le modèle

$$Y_i = \mu_0 + \beta_{1l} G_{i1l} + \cdots + \beta_{pl} G_{ipl} + \epsilon_i, \tag{3.3}$$

si le phénotype continu ou bien le modèle

$$\text{logit}\{P(Y_i = 1 \mid G_{i1l}, \ldots, G_{ipl})\} = \mu_0 + \beta_{1l} G_{i1l} + \cdots + \beta_{pl} G_{ipl}, \tag{3.4}$$

si le phénotype est binaire. Par la suite, on procède à un test d'hypothèse global de la région l qui se traduit formellement par :

$$\mathrm{H}_0 : \boldsymbol{\beta}_l = 0_p \quad \text{contre} \quad \mathrm{H}_a : \exists\, j \in \{1, \ldots, p\} : \beta_{jl} \neq 0,$$

en utilisant un test de rapport de vraisemblance ou un test de Wald.
Dans les études d'association sur tout le génome (GWAS), on considère souvent un très grand nombre q de régions chromosomiques. Ainsi, lorsqu'on effectue un test pour chaque $\boldsymbol{\beta}_l$ où $l \in \{1, \ldots, q\}$, le seuil global, à savoir la probabilité de rejeter H_0 pour au moins une région alors que celle ci est vraie (H_0), sera beaucoup plus grand que α. Afin de contrôler le seuil global, on utilise en général la correction de Bonferroni qui suggère de réduire le seuil de chaque test à (α/q) où q est le nombre total de régions que l'on veut tester.

3.4 Contrôle de la structure de population

Les approches décrites jusqu'ici reposent sur deux hypothèses fondamentales. Tout d'abord, les individus qui rentrent dans l'étude sont supposés provenir d'une seule population génétiquement homogène. Autrement dit, il ne devrait y avoir aucune structure dans la population (on parle aussi de stratification de la population dans la littérature). La stratification apparaît lorsque les individus de l'échantillon ont des ancêtres distincts en remontant plus loin dans la généalogie.
Deuxièmement, tous les individus de l'échantillon doivent représenter des unités statistiquement indépendantes tirées de cette même population homogène, et cela

n'est pas toujours vrai. En effet, en réalité, même si les individus génotypés ne proviennent pas en général de la même famille (ce qui n'est toutefois pas impossible), ils partagent néanmoins des ancêtres communs mais de façon plus lointaine dans le temps, c'est ce que l'on appelle "la parenté cachée" ou *cryptic relatedness* en anglais.

Ainsi, si l'une des deux suppositions précédentes n'est pas respectée, les tests d'association standard ont tendance à avoir une inflation de l'erreur de type 1.

L'erreur de type 1 se traduit par le fait de rejeter à tort l'hypothèse d'absence d'association, autrement dit, conclure à une association avec le phénotype alors que celle ci n'est pas vraie (un faux positif). Cela se produit en général lorsque les fréquences d'allèles des individus échantillonnés varient entre les sous-populations. Un ou plusieurs de ces allèles peuvent en effet être impliqués dans la détermination du phénotype, mais les statistiques standard peuvent ne pas les distinguer des nombreux allèles à l'échelle du génome et dont les fréquences varient selon les sous-populations à cause de la dérive génétique ou de la sélection naturelle (voir Astle et Balding, 2009). Pour donner un exemple : prenons le cas d'un grand échantillon composé de Chinois et de Canadiens. De nombreux variants génétiques sont susceptibles de présenter une association avec le phénotype "savoir manger avec des baguettes". Ceux ci seront les allèles qui sont relativement communs au Chinois qui ont une histoire différente des Canadiens mais qui ne causent pas le fait de savoir manger avec des baguettes (gène de la baguette !).

Pour remédier à ce problème, diverses approches ont été développées dans la littérature et nous allons décrire les plus utilisées.

3.4.1 Le contrôle génomique

Le contrôle génomique est une technique simple pour réduire l'inflation de l'erreur de type 1 causée par la structure de la population ou la *cryptic relatedness*. Elle

consiste à corriger la statistique du test par un facteur d'inflation constant noté λ. Ce facteur est calculé sur la base des SNPs nuls (qui montrent une absence d'association avec le phénotype) de la façon suivante :

$$\lambda = \frac{\text{Médiane } (T_1, \ldots, T_k)}{M},$$

où $T_1, \ldots, T_k$ sont les valeurs de la statistique du test des k marqueurs non associés et M la médiane théorique de la statistique du test. En fait, l'idée est que si la valeur médiane des SNPs nuls s'éloigne de la valeur théorique ($\lambda > 1$), alors cela est certainement dû à la présence d'une structure dans la population. Tandis qu'une valeur de $\lambda \approx 1$ indique une absence de structure. En cas de présence de structure, il suffit alors de diviser les statistiques du test par λ pour contrôler l'effet de la structure.

Toutefois, selon Price *et al.* (2010), il faut distinguer entre les différences des sous-populations qui sont dues à une récente dérive génétique de celles qui proviennent d'une ancienne division de la population. Dans le premier cas, une correction en divisant la statistique du test par λ peut suffire tandis que dans le second cas, cela peut s'avérer insuffisant en raison de la présence de marqueurs avec des différences inhabituelles dans la fréquence d'allèles probablement causées par la sélection naturelle.

D'autres approches de correction de la stratification, y compris des approches qui tiennent compte également de la parenté cachée (cryptic relatedness), sont décrits ci-dessous.

3.4.2 L'analyse en composantes principales (ACP)

Afin de contrôler l'effet de la structure de population, Zhang *et al.* (2003) proposent d'inclure les composantes principales de la matrice de génotype $\boldsymbol{G}$ comme covariables dans le modèle de régression (3.3) décrit précédemment. Price *et al.* (2006) présentent la même méthode mais appliquée à des études cas-témoins.

L'idée est d'utiliser les marqueurs nuls (non-associés avec le phénotype) pour construire les composantes principales qui décrivent de la meilleure façon la variabilité entre individus. Soit $\boldsymbol{G}$ la matrice de génotype constituée de n lignes correspondant aux individus et de p colonnes correspondant aux marqueurs génétiques. Les génotypes sont codés de façon à compter le nombre d'allèles mineurs pour chaque individu, autrement dit, $G_{ij} \in \{0, 1, 2\}$, mais sont standardisés par rapport aux colonnes de façon à avoir une moyenne nulle et une variance égale à 1. Étant donné que se sont les individus qui nous intéressent, l'ACP est réalisée sur la matrice $\boldsymbol{\Phi} = (1/p)\boldsymbol{G}\boldsymbol{G}^{\boldsymbol{T}}$ plutôt que sur la matrice des corrélations de $\boldsymbol{G}$. La matrice $\boldsymbol{\Phi}$ aussi appelée matrice *kinship* empirique permet de tenir compte de la similarité génotypique entre les individus. En effet, l'élément ϕ_{ij} de $\boldsymbol{\Phi}$ est une mesure empirique de la proportion du génome au complet que les individus (i, j) partagent par descendance. Les composantes principales sont obtenues à partir de la décomposition spectrale de la matrice $\boldsymbol{\Phi}$ donnée par :

$$\boldsymbol{\Phi} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\boldsymbol{T}},$$

où $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_n)$ est une matrice diagonale contenant les valeurs propres de $\boldsymbol{\Phi}$ et $\boldsymbol{U} = [\boldsymbol{u_1}; \ldots; \boldsymbol{u_n}]$ est la matrice des vecteurs propres associés. Ainsi, les composantes principales auront une forte corrélation avec les SNPs dont les fréquences d'allèles varient entre les sous populations. De façon générale, pour un modèle en k sous-populations, il suffit d'inclure les $(k-1)$ premières composantes principales pour corriger l'effet la structure (Price *et al.*, 2006).

3.4.3 Modèle linéaire mixte

Une autre façon de traiter le problème de la confusion lié à la structure de population et à la *cryptic relatedness* est de considérer un modèle linéaire mixte. Ce modèle intègre à la fois des effets fixes (les SNPs candidats et autres covariables telles que le sexe, l'âge, ... etc), comme dans le modèle linéaire standard, mais aussi

un effet aléatoire qui permet de capturer une partie de la variation phénotypique due à la structure de population. Autrement dit, la partie résiduelle du modèle de régression (3.3) en présence de structure dans la population peut se décomposer en deux parties comme suit :

$$\epsilon = \delta + u,$$

où δ est un effet aléatoire qui modélise la variation résiduelle génétique sur le phénotype (effets héritables non-observables) et dont la structure de variance-covariance dépend de la matrice Φ décrite précédemment, tandis que u représente l'effet non-héritable et non-observable ou autrement dit, la nouvelle partie résiduelle. On peut finalement écrire ce modèle sous la forme matricielle suivante :

$$Y = \underbrace{X\alpha + G\beta}_{\text{effet fixe}} + \underbrace{\delta}_{\text{effet aléatoire}} + \underbrace{u}_{\text{partie résiduelle}},$$

où $Y_{(n\times 1)}$ représente le vecteur du phénotype supposé normalement distribué, $X_{(n\times m)}$ est la matrice de covariables (âge, sexe,...) et $\alpha_{(m\times 1)}$ le vecteur d'effets associés aux covariables. La matrice $G_{(n\times p)}$ contient les génotype des individus et $\beta_{(p\times 1)}$ l'effet correspondant sur le phénotype. Enfin, $\delta_{(n\times 1)}$ est un vecteur d'effets aléatoires supposé normal de moyenne nulle et de variance $\sigma^2_\delta \Phi_{(n\times n)}$. Ainsi,

$$\text{Var}(Y) = \text{Var}(\delta) + \text{Var}(u) = \sigma^2_\delta \Phi + \sigma^2_u I_n.$$

Les méthodes d'association que nous avons décrites jusqu'ici ont permis l'identification de nombreux loci associés à des maladies complexes humaines (Hindorff *et al.*, 2009). Cependant, les variants génétiques communs identifiés n'expliquent qu'une faible proportion du trait (ou du caractère) héritable. Pour pallier à ce problème de manque d'héritabilité, les chercheurs postulent une autre hypothèse selon laquelle les variants génétiques rares dont la fréquence allélique est inférieure à 5% dans la population ont une grande influence sur les maladies (Schork *et al.*, 2009). Toutefois, la puissance des tests sur les variants rares avec les méthodes

standard s'est révélée inférieure à la puissance pour les variants communs sauf dans le cas où l'on dispose d'une taille d'échantillon très grande (Li et Leal, 2008). C'est pourquoi, le développement de nouveaux tests pour variants rares a suscité un grand intérêt ces dernières années. Nous allons donc présenter quelques uns de ces tests à la section suivante.

3.5 Tests d'association sur les variants rares

Après l'échec des méthodes standard face aux variants rares, de nombreux tests ont été proposés dans la littérature. Cette section sera consacrée à la présentation de quelques uns de ces tests de façon plus ou moins détaillée.

3.5.1 Les tests d'association "burden"

Les tests d'association *"burden"* (BT par la suite) sont apparus comme première alternative aux méthodes standards. Ils permettent de résumer l'information des variants rares d'une région en une seule variable qui permettra par la suite de tester les effets cumulés de ces variants sur le phénotype. Le modèle de régression généralisé pour les *burden tests* s'écrit comme suit :

$$g\{\mathbb{E}(Y_i)\} = \boldsymbol{X}_i^T\boldsymbol{\alpha} + \beta_c \sum_{j=1}^{p} w_j G_{ij},$$

où $g(\cdot)$ est la fonction lien et $\boldsymbol{\alpha}$ et β_c sont les coefficients de régression de $\boldsymbol{X_i}$ et $\sum_{j=1}^{p} w_j G_{ij}$ respectivement. Le coefficient w_j correspond au poids du $j^{\text{ème}}$ variant génétique et peut être fixé en fonction de la fréquence de l'allèle mineur du variant par exemple. L'association entre le phénotype et la région est testée par $\mathrm{H}_0 : \beta_c = 0$ en utilisant un test de score. La statistique du test est donnée par :

$$Q_{BT} = \left[\sum_{i=1}^{n}(Y_i - \hat{Y}_i)\left(\sum_{j=1}^{p} w_j G_{ij}\right)\right]^2,$$

qui suit approximativement une loi de khi-deux à un degré de liberté sous l'hypothèse nulle. Ces tests supposent implicitement que tous les variants sont causaux

et ont le même effet sur le phénotype. Ainsi, si ces hypothèses sont vérifiées, les BT auront une bonne puissance. En revanche, la violation de ces hypothèses peut conduire à une perte de puissance.

3.5.2 Le test d'association SKAT

Après les *burden tests*, une autre catégorie de tests connue sous le nom des tests sur la composante de variance (*variance component tests*) est apparue.
SKAT pour *Sequence kernel association test* (Wu *et al.*, 2011) est un modèle linéaire mixte qui permet de tester l'effet combiné de plusieurs variants génétiques (rares ou communs) sur un phénotype. Ce dernier peut être continu ou binaire comme dans les études cas-témoins. Le fait de travailler avec des variants rares peut rendre impossible l'estimation de tous les effets β_j. Pour pallier à ce problème, SKAT suppose que les effets liés à la composante génétique sont complètement aléatoires de moyenne nulle et de variance $\tau \cdot w_j$ où w_j désigne le poids spécifique à chaque variant génétique et τ la composante de variance. La forme matricielle du modèle pour un trait quantitatif $\boldsymbol{Y}$ s'écrit alors comme suit :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.5}$$

où $\boldsymbol{X}$ désigne la matrice de covariables et $\boldsymbol{\alpha}$ le vecteur des effets associés. La matrice $\boldsymbol{G}$ contient les génotypes des individus d'une région constituée de p marqueurs et $\boldsymbol{\beta}$ le vecteur des effets de la composante génétique avec $\boldsymbol{\beta} \sim N(0_p, \tau \boldsymbol{W})$ et $\boldsymbol{\epsilon} \sim N(0_n, \sigma_\epsilon^2 \boldsymbol{I}_n)$. Ainsi,

$$\begin{aligned} \mathbb{E}(\boldsymbol{Y}) &= \boldsymbol{X}\boldsymbol{\alpha}, \\ \mathrm{Var}(\boldsymbol{Y}) &= \tau \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}^{\boldsymbol{T}} + \sigma_\epsilon^2 \boldsymbol{I}_n. \end{aligned}$$

Pour tester l'hypothèse d'association, les auteurs de SKAT proposent le test

$$\mathrm{H}_0 : \tau = 0 \quad \text{contre} \quad \mathrm{H}_a : \tau > 0,$$

au lieu du test

$$\mathrm{H}_0 : \boldsymbol{\beta} = 0_p \quad \text{contre} \quad \mathrm{H}_a : \boldsymbol{\beta} \neq 0_p.$$

En effet, puisque l'on suppose que les effets s'annulent en moyenne, et si leur variance est nulle alors cela veut dire que $\beta_i = \beta_j = 0$ pour tout $i, j \in \{1, \ldots, p\}$. Pour réaliser ce test, SKAT utilise un test de score au lieu d'un test de rapport de vraisemblance ou un test de Wald comme auparavant. Ainsi, SKAT nécessite l'estimation des paramètres du modèle sous l'hypothèse nulle seulement, ce qui le rend très avantageux computationnellement. La statistique du test est donnée par :

$$Q = (\boldsymbol{Y} - \hat{\boldsymbol{\mu}})^{\boldsymbol{T}} \boldsymbol{K} (\boldsymbol{Y} - \hat{\boldsymbol{\mu}}),$$

où $\hat{\boldsymbol{\mu}} = \boldsymbol{X}\hat{\boldsymbol{\alpha}}$ représente la moyenne prédite de $\boldsymbol{Y}$ sous H_0. La loi de Q sous l'hypothèse nulle est un mélange de loi de khi-deux. La façon d'obtenir la statistique du test n'est pas présenté dans cette section mais est très similaire à celle que nous présenterons en détail au chapitre IV pour notre modèle.

La matrice $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}^{\boldsymbol{T}}$ aussi appelée noyau linéaire pondéré (ou *linear weighted kernel*), permet de tenir compte de la similarité génotypique entre individus (deux à deux) dans une région chromosomique. La matrice diagonale $\boldsymbol{W}$ de dimension $(p \times p)$ contient les poids associés aux variants génétiques. Il est important de noter qu'un bon choix des poids w_j permet d'améliorer la puissance. Une valeur de w_j proche de zéro, donnera au $j^{\text{ème}}$ marqueur une petite contribution à Q. Ainsi, donner plus de poids aux variants causaux et diminuer celui des non-causaux peut améliorer significativement la puissance. Cependant, puisque l'on ignore quels sont les variants causaux, les auteurs de SKAT proposent de fixer les poids à l'aide de la loi béta de sorte que $\sqrt{w_j} = \mathrm{Beta}(MAF_j; a_1 = 1, a_2 = 25)$ où MAF_j représente la fréquence de l'allèle mineur du $j^{\text{ème}}$ variant génétique, afin d'accroître le poids des variants rares dont la fréquence d'allèles varient entre 1% et 5%. Lorsqu'on fixe $a_1 = a_2 = 1$, cela correspond à $w_j = 1$. Autrement dit, donner à tous les

variants le même poids.

SKAT montre une bonne puissance quand une grande partie des variants de la région sont non causaux ou que les effets des variants causaux ont des directions différentes. Par contre, lorsque la majorité des variants sont causaux et ont tous des effets dans la même direction les BT sont plus avantageux que SKAT.

SKAT-O pour *optimal* SKAT (Lee *et al.*, 2012) est un test d'association combinant à la fois SKAT et les *burden test* de façon à maximiser la puissance du test. Ce test utilise la statistique :

$$Q_{Opt}(\rho) = \rho Q_{BT} + (1-\rho)Q_{SKAT},$$

où Q_{BT} et Q_{SKAT} sont les statistiques des BT et de SKAT décrites précédemment. La valeur de $\rho \in [0,1]$, peut être interprétée comme la corrélation des coefficients de régression β_j où $j \in \{1,\ldots,p\}$. Lorsque ces coefficients sont parfaitement corrélés ($\rho = 1$), la statistique $Q_{Opt}(1) = Q_{BT}$. Cependant, puisque l'on ignore la véritable valeur de ρ, les auteurs de SKAT-O proposent de calculer $Q_{Opt}(\rho)$ pour différentes valeurs de ρ tel que $0 = \rho_1 < \rho_2 < \cdots < \rho_b = 1$, et de retenir celle qui maximise la puissance. Nous invitons le lecteur désirant plus de détails à consulter l'article de Lee *et al.* (2012).

3.5.3 Le test de score "MiST"

Mixed effect score test ou MiST (Sun *et al.*, 2013) est aussi un test d'association basé sur un modèle linéaire mixte comme SKAT sauf qu'il permet d'inclure de l'information additionnelle concernant les variants génétiques en décomposant l'effet de la composante génétique en deux parties. La première partie permet de tenir compte de l'effet individuel de chaque variant sur le phénotype comme SKAT tandis que la deuxième partie permet de supposer que les variants ayant les mêmes caractéristiques ont le même effet sur le phénotype. Le test est réalisé en combinant deux tests de scores.

CHAPITRE IV

INFORMATION GÉNÉTIQUE ET GÉNÉALOGIQUE AU SERVICE DES TESTS D'ASSOCIATION

Nous avons présenté au chapitre précédent différentes méthodes de cartographie génétique qui permettent de localiser la position d'un variant génétique associé à un phénotype d'intérêt. Pour ce qui est des méthodes d'association, nous avons vu que celles-ci pouvaient conduire à de fausses associations dans le cas où une structure de population est présente dans notre échantillon.

Dans ce chapitre, nous proposons un nouveau modèle pour tester s'il y a ou non une association du trait avec la région chromosomique considérée. Il s'agit d'un modèle linéaire mixte tel que décrit dans l'équation (3.5) mais dans lequel nous avons ajouté un effet aléatoire qui permet de tenir compte de la corrélation entre individus afin de contrôler la structure de population. Cette idée rejoint celle de Oualkacha *et al.* (2013) qui a été appliquée à des données familiales. La structure de variance de ce nouveau terme aléatoire dépend d'une nouvelle matrice de similarité notée $\boldsymbol{S_{TMRCA}}$ construite en utilisant le processus de coalescence avec recombinaison. L'idée est en fait simple : la similarité entre individus est mesurée par le temps de coalescence entre chaque paire d'individus de notre échantillon. Ainsi, lorsque le temps de coalescence est court, cela indique une similarité (ou une ressemblance) au niveau de leur caractéristiques génétiques. À l'opposé, lorsque le temps de coalescence est grand, cela reflète plutôt une dissimilarité.

Nous allons à présent nous pencher sur la façon d'obtenir $\boldsymbol{S_{TMRCA}}$. Par la suite, nous développerons en détail la façon de réaliser le test d'association.

4.1 Construction de la matrice S_{TMRCA}

Comme nous l'avons déjà mentionné, la matrice $\boldsymbol{S_{TMRCA}}$ est une mesure de similarité génotypique basée sur le temps à l'ancêtre commun. En fait, $\boldsymbol{S_{TMRCA}}$ est le résultat d'une transformation (voir section 4.1.4) sur une matrice symétrique de dimension $(2n \times 2n)$ notée $\boldsymbol{T^{hap}}$ dont les éléments représentent le temps de coalescence entre chaque paire d'haplotypes des individus de notre échantillon. La forme de $\boldsymbol{T^{hap}}$ est donnée comme suit :

$$\boldsymbol{T^{hap}} = \begin{bmatrix} 0 & T_{1,2} & T_{1,3} & \cdots & T_{1,2n} \\ T_{1,2} & 0 & T_{2,3} & \cdots & T_{2,2n} \\ T_{1,3} & T_{2,3} & 0 & \cdots & T_{3,2n} \\ \vdots & \vdots & \vdots & \ddots & \\ T_{1,2n} & T_{2,2n} & T_{3,2n} & \cdots & 0 \end{bmatrix}, \tag{4.1}$$

où $T_{i,j}$ représente le temps de coalescence entre les haplotypes i et j de notre échantillon avec $i, j \in \{1, 2, \ldots, 2n\}$. Il est à noter que le temps de coalescence d'un haplotype avec lui même $T_{i,i} = 0$ par définition.
Nous allons dans un premier temps décrire la façon d'obtenir $\boldsymbol{T^{hap}}$, puis exposer les détails de la transformation pour obtenir $\boldsymbol{S_{TMRCA}}$.

4.1.1 Retour sur le processus de coalescence

Pour obtenir la matrice $\boldsymbol{T^{hap}}$, nous devons reconstruire la généalogie (ou l'ARG) qui relie les individus de notre échantillon à leur plus récent ancêtre commun. Cependant, il est impossible de connaître avec exactitude la véritable généalogie ou le déroulement exact de l'histoire qui a permis d'obtenir cet échantillon. Toutefois, nous pouvons construire diverses généalogies (probables) afin d'estimer un temps

de coalescence moyen. Ce temps est donné par :

$$\mathbb{E}(T_{MRCA}) = \int t_{mrca}(G) \cdot P(H_0 \mid G) \cdot P(G) \, \mathrm{dG}, \tag{4.2}$$

où H_0 désigne l'ensemble des haplotypes de notre échantillon (où chaque individu est représenté par deux haplotypes pour tenir compte de la réalité diploïde), G est un ARG, $t_{mrca}(G)$ est le temps à l'ancêtre commun pour la généalogie G et $P(H_0 \mid G)$ est une fonction indicatrice qui représente la probabilité qu'une certaine généalogie G donne l'ensemble des haplotypes observés de l'échantillon H_0. Ainsi, si $P(H_0 \mid G) = 1$, on dira que la généalogie G est cohérente avec les données de l'échantillon. Une façon d'avoir des généalogies cohérentes est de les construire à l'aide d'une chaîne de Markov dont l'état initial est H_0. En utilisant uniquement les généalogies cohérentes (autrement dit $P(H_0|G) = 1$), l'équation (4.2) devient :

$$\mathbb{E}(T_{MRCA}) = \int t_{mrca}(G) \cdot P(G) \, \mathrm{dG}. \tag{4.3}$$

Cependant, à cause des événements de recombinaison, l'espace des ARG est infini, ce qui rend l'équation (4.3) impossible à évaluer. La solution consiste alors à générer un ensemble fini de graphe en utilisant l'échantillonnage préférentiel (*Importance sampling*) en accordant plus de poids à certaines généalogies qu'à d'autres. On obtient alors :

$$\mathbb{E}(T_{MRCA}) \;=\; \int t_{mrca}(G) \cdot \frac{P(G)}{Q(G)} \cdot Q(G) \, \mathrm{dG}. \tag{4.4}$$

Afin d'évaluer l'équation (4.4), on peut utiliser une approximation Monte Carlo, donnée par l'équation (4.5) ci dessous :

$$\bar{T}_{MRCA} = \widehat{\mathbb{E}(T_{MRCA})} = \frac{1}{M} \sum_{i=1}^{M} \frac{P(G^{(i)})}{Q(G^{(i)})} \cdot t_{mrca}(G^{(i)}), \tag{4.5}$$

où $G^{(1)}, G^{(2)}, \ldots, G^{(M)}$ sont générés à partir de la distribution Q.

Les sections 4.1.1 et 4.1.2 traiteront de la façon de calculer la probabilité d'un graphe de recombinaison ancestral $P(G)$ ainsi que la distribution proposée Q de Fearnhead et Donnelly (2001) pour générer des généalogies.

4.1.2 Probabilité d'un graphe de recombinaison ancestral

Avant d'entrer dans les détails, nous allons tout d'abord introduire quelques notations. Soit H_τ la configuration de l'échantillon à l'étape τ de la construction du graphe. En effet, la construction d'un graphe nécessite le passage de l'état H_0 composé de $2n$ haplotypes à l'état H_{τ^*} composé d'un seul haplotype (haplotype ancestral) en passant par les états successifs $H_1, H_2, \ldots, H_{\tau^*-1}$, où chaque état est le résultat d'un événement de coalescence, de mutation ou de recombinaison sur l'état précédent. Ainsi, il s'agit d'un processus markovien dont la probabilité d'un état ne dépend du passé que par l'état précèdent. On obtient alors :

$$\begin{aligned}
P(G) &= P(H_0, H_1, \ldots, H_{\tau^*}) \\
&= P(H_0 \mid H_1, \cdots, H_{\tau^*}) \cdot P(H_1, H_2, \cdots, H_{\tau^*}) \\
&= P(H_0 \mid H_1) \cdot P(H_1, H_2, \cdots, H_{\tau^*}) \\
&= P(H_0 \mid H_1) \cdot P(H_1 \mid H_2) \cdot P(H_2, H_3, \cdots, H_{\tau^*}) \\
&\vdots \\
&= P(H_{\tau^*}) \cdot \prod_{\tau=0}^{\tau^*-1} P(H_\tau \mid H_{\tau+1}) \qquad (4.6) \\
&= \prod_{\tau=0}^{\tau^*-1} P(H_\tau \mid H_{\tau+1}). \qquad (4.7)
\end{aligned}$$

Le passage de l'équation (4.6) à l'équation (4.7) se fait sous l'hypothèse que l'haplotype ancestral est connu (allèle 0 à tous les marqueurs), ainsi, $P(H_{\tau^*}) = 1$.

À présent, intéressons nous à la façon de calculer la probabilité conditionnelle $P(H_\tau \mid H_{\tau+1})$. On se souvient que dans la construction d'un ARG, chaque étape est soit le résultat d'une coalescence, d'une mutation ou d'une recombinaison.

Soit $n = (n_1, n_2, \ldots, n_k)$ la configuration de l'échantillon à une étape τ du graphe. Nous observons ainsi, k types de séquences où la séquence de type i est de multiplicité n_i. Une coalescence peut se produire soit entre deux séquences du même type i, on notera cet événement par C_i, soit entre deux séquences de type i et

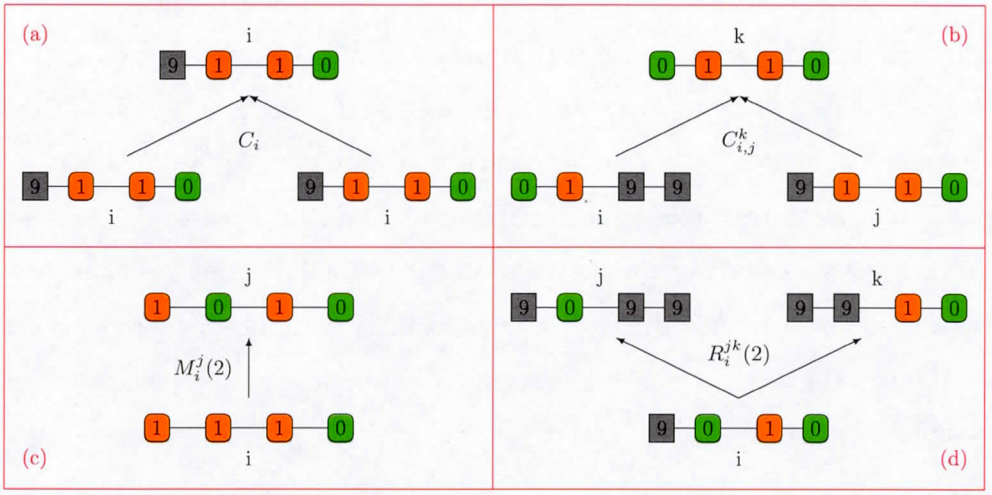


Figure 4.1 Exemple des différents types d'événements possibles, où les carrés gris représentent des marqueurs non-ancestraux. La figure (a) représente une coalescence entre deux séquences identiques, la figure (b) illustre une coalescence entre deux séquences différentes, un événement de mutation au marqueur 2 est illustré à la figure (c) et enfin la figure (d) représente une recombinaison ayant lieu entre le marqueur 2 et 3 d'une séquence de type i.

j qui différent uniquement par le matériel non-ancestral (dû aux événements de recombinaison) et résultant en une séquence de type k. Cet événement sera noté $C^k_{i,j}$. Les figures 4.1.a et 4.1.b respectivement, illustrent ces deux événements de coalescence.

Une mutation peut avoir lieu lorsqu'il reste dans l'échantillon une seule séquence de type i portant l'allèle mutant (1) au marqueur m (sous l'hypothèse du modèle de sites infinis) et résultant en une séquence de type j. On note cet événement par $M^j_i(m)$ tel qu'illustré sur la figure 4.1.c.

Un événement de recombinaison peut se produire uniquement si il existe au moins un marqueur ancestral (mutant ou non) de chaque coté du point de recombinaison

s car les séquences constituées uniquement de matière non ancestrale sont non informatives pour notre graphe. Ainsi, on notera par $R_i^{jk}(s)$ la recombinaison d'une séquence de type i en deux séquences de type j et k comme l'illustre la figure 4.1.d. Après chaque étape, lors de la construction d'un graphe, la probabilité des trois événements décrits précédemment doit être mis à jour à cause de la présence de matière non-ancestrale due aux événements de recombinaison.

Soit α la proportion de marqueurs ancestraux parmi l'ensemble de marqueurs présents à un certain état H_τ et β la proportion des fragments de séquences où une recombinaison peut avoir lieu. Ainsi :

$$\alpha = \frac{\sum_i n_i \cdot a_i}{np} \quad \text{et} \quad \beta = \frac{\sum_i n_i \cdot r_i}{nr},$$

où :

- a_i est le nombre de marqueurs ancestraux contenus dans une séquence de type i ;
- p est le nombre total de marqueurs d'une séquence ;
- r_i est la distance entre les marqueurs ancestraux situés aux extrémités d'une séquence de type i ;
- r est la longueur totale d'une séquence.

Les trois probabilités associées aux trois événements précédents deviennent :

$$\begin{aligned} P_\tau(C) &= \frac{n-1}{n-1+\alpha\theta+\beta\rho}, \\ P_\tau(M) &= \frac{\alpha\theta}{n-1+\alpha\theta+\beta\rho}, \\ P_\tau(R) &= \frac{\beta\rho}{n-1+\alpha\theta+\beta\rho}. \end{aligned}$$

Maintenant, nous sommes en mesure d'obtenir la distribution de $P(H_\tau \mid H_{\tau+1})$. Premièrement, si une coalescence entre deux séquences non-identiques a lieu entre les états H_τ et $H_{\tau+1}$, alors on peut écrire $H_{\tau+1} = H_\tau + C_{i,j}^k$. Un tel événement implique une diminution du nombre total de séquences de un à l'état $H_{\tau+1}$. Autrement dit, si $\text{card}(H_\tau) = n$ alors $\text{card}(H_{\tau+1}) = n - 1$. Ainsi, la probabilité

conditionnelle d'intérêt est obtenue en multipliant la probabilité d'une coalescence entre les états H_τ et $H_{\tau+1}$ par la probabilité qu'une séquence de type k choisie au hasard parmi les $(n_k + 1 - \delta_{ik} - \delta_{jk})$ séquences à l'état $H_{\tau+1}$ soit le résultat de cette coalescence. Notons que $\delta_{rs} = 1$ si $r = s$ et 0 sinon. Ce qui nous permet d'écrire :

$$P(H_\tau \mid H_{\tau+1} = H_\tau + C^k_{i,j}) = P_\tau(C) \cdot \frac{n_k + 1 - \delta_{ik} - \delta_{jk}}{n-1}.$$

Lorsque nous avons une coalescence entre deux séquences identiques (disons de type i), autrement dit $H_{\tau+1} = H_\tau + C^i_{i,i} = H_\tau + C_i$, la probabilité conditionnelle devient :

$$\begin{aligned} P(H_\tau \mid H_{\tau+1} = H_\tau + C_i) &= P_\tau(C) \cdot \frac{n_i + 1 - \delta_{ii} - \delta_{ii}}{n-1} \\ &= P_\tau(C) \cdot \frac{n_i + 1 - 1 - 1}{n-1} \\ &= P_\tau(C) \cdot \frac{n_i - 1}{n-1}. \end{aligned}$$

Deuxièmement, si une mutation a lieu au marqueur m (marqueur ancestral) d'une séquence de type i, on aura donc une séquence de type j en plus à l'état $H_{\tau+1}$ pour remplacer la séquence de type i. Ainsi, la probabilité qu'une séquence de type j choisie au hasard parmi les n séquences (un événement de mutation ne modifie pas le nombre de séquences) présentes à l'état $H_{\tau+1}$ mute au marqueur m est de $(n_j + 1/n) \cdot (1/\alpha p)$ où p est le nombre total de marqueurs et $(1/\alpha p)$ représente la probabilité d'avoir une mutation au marqueur m. Nous obtenons donc :

$$P(H_\tau \mid H_{\tau+1} = H_\tau + M^j_i(m)) = P_\tau(M) \cdot \frac{n_j + 1}{n} \cdot \frac{1}{\alpha p}.$$

Enfin, si une séquence de type i recombine en séquences de type j et k. Alors, les nombres respectifs des ces séquences passent de n_j à $(n_j + 1)$ et de n_k à $(n_k + 1)$ à l'étape $H_{\tau+1}$. D'où :

$$P(H_\tau \mid H_{\tau+1} = H_\tau + R^{j,k}_i(s)) = P_\tau(R) \cdot \frac{(n_j + 1)(n_k + 1)}{n(n+1)} \cdot \frac{r_s}{\beta r},$$

où $(r_s/\beta r)$ est la probabilité qu'une recombinaison ait lieu dans l'intervalle s.

4.1.3 Distribution de Fearnhead et Donnelly

Nous avons mentionné précédemment que les ARGs sont générés à l'aide d'une distribution proposée Q. Plusieurs distributions ont été proposée dans la littérature notamment celle de Griffiths et Marjoram (1996) et de Fearnhead et Donnelly (2001). Cette dernière a été implémenté par Descary (2012) dans le programme TMRCA que nous avons utilisé dans nos simulations pour obtenir la matrice $\boldsymbol{T^{hap}}$. Nous allons donc nous intéresser uniquement à cette distribution. $Q(G)$ peut être calculée de façon similaire que $P(G)$. Ainsi, nous avons :

$$\begin{aligned}
Q(G) &= Q(H_0, H_1, \cdots, H_{\tau^*}) \\
&= Q(H_{\tau^*} | H_0, H_1, \cdots, H_{\tau^*-1}) \cdot Q(H_0, H_1, \cdots, H_{\tau^*-1}) \\
&= Q(H_{\tau^*} \mid H_{\tau^*-1}) \cdot Q(H_{\tau^*-1} \mid H_0, H_1, \cdots, H_{\tau^*-2}) \cdot Q(H_0, H_1, \cdots, H_{\tau^*-2}) \\
&= Q(H_{\tau^*} \mid H_{\tau^*-1}) \cdot Q(H_{\tau^*-1} \mid H_{\tau^*-2}) \cdot Q(H_0, H_1, \cdots, H_{\tau^*-3}) \\
&\vdots \\
&= Q(H_0) \cdot \prod_{\tau=0}^{\tau^*-1} Q(H_{\tau+1} \mid H_\tau) \qquad (4.8) \\
&= \prod_{\tau=0}^{\tau^*-1} Q(H_{\tau+1} \mid H_\tau). \qquad (4.9)
\end{aligned}$$

Le passage de l'équation (4.8) à l'équation (4.9) se base sur le fait que l'état initial de la chaîne de Markov que l'on utilise pour construire les graphe est H_0 (ce qui nous permet, entre autres, d'avoir des généalogies cohérentes). Ainsi, $Q(H_0) = 1$. Fearnhead et Donnelly ont démontré que la distribution optimale a pour probabilité de transition :

$$Q(H_{\tau+1} \mid H_\tau) = P(H_\tau | H_{\tau+1}) \cdot \frac{\phi(H_{\tau+1})}{\phi(H_\tau)}, \qquad (4.10)$$

où $\phi(H_\tau)$ désigne la probabilité qu'un échantillon de séquences tiré au hasard d'une population soit identique à celui de H_τ en considérant uniquement le matériel ancestral.

En remplaçant les équations (4.7) et (4.10) dans l'équation (4.5), nous obtenons :

$$\bar{T}_{MRCA} = \mathbb{E}(\widehat{T_{MRCA}}) = \frac{1}{M}\sum_{i=1}^{M}\left(\prod_{\tau=0}^{\tau^*-1}\frac{\phi(H_\tau^{(i)})}{\phi(H_{\tau+1}^{(i)})}\right)\cdot t_{mrca}(G^{(i)}).$$

La distribution $\phi(H_\tau)$ demeure toutefois inconnue et il faut donc l'estimer. Pour cela, on définit $\phi(\gamma \mid H_\tau - \gamma)$, la probabilité conditionnelle de tirer aléatoirement d'une population, une séquence de type γ pour compléter l'échantillon H_τ alors que l'on a déjà $(\text{card}(H_\tau) - 1)$ séquences. Ainsi, γ représente le type de la dernière séquence tirée pour compléter l'échantillon H_τ. Nous pouvons donc écrire :

$$\phi(H_\tau) = \phi(\gamma \mid H_\tau - \gamma)\cdot\phi(H_\tau - \gamma). \tag{4.11}$$

En se servant de l'équation (4.11), nous pouvons écrire le rapport $\phi(H_\tau)/\phi(H_{\tau+1})$ en fonction du type de la dernière séquence tirée pour les trois événements possibles (coalescence, mutation et recombinaison).

Premièrement, si $H_{\tau+1} = H_\tau + C_{i,j}^k \equiv H_\tau - i - j + k$, alors nous avons :

$$\begin{aligned}\frac{\phi(H_\tau)}{\phi(H_{\tau+1})} &= \frac{\phi(j \mid H_\tau - i - j)\phi(H_\tau - i - j)\phi(i \mid H_\tau - i)}{\phi(H_\tau - i - j + k)}\\ &= \frac{\phi(j \mid H_\tau - i - j)\phi(H_\tau - i - j)\phi(i \mid H_\tau - i)}{\phi(k \mid H_\tau - i - j)\phi(H_\tau - i - j)}\\ &= \frac{\phi(j \mid H_\tau - i - j)\phi(i \mid H_\tau - i)}{\phi(k \mid H_\tau - i - j)}.\end{aligned}$$

Dans le cas où nous avons une coalescence entre deux séquences identiques, nous obtenons :

$$\frac{\phi(H_\tau)}{\phi(H_{\tau+1})} = \frac{\phi(i|H_\tau - i - i)\phi(i|H_\tau - i)}{\phi(i|H_\tau - i - i)} = \phi(i|H_\tau - i).$$

Deuxièmement, si nous avons un événement de mutation, alors une séquence de type i est remplacée par une séquence de type j. Autrement dit, $H_{\tau+1} = H_\tau + M_i^j \equiv H_\tau - i + j$. Nous aurons alors :

$$\frac{\phi(H_\tau)}{\phi(H_{\tau+1})} = \frac{\phi(i \mid H_\tau - i)\phi(H_\tau - i)}{\phi(H_\tau - i + j)} = \frac{\phi(i \mid H_\tau - i)\phi(H_\tau - i)}{\phi(j \mid H_\tau - i)\phi(H_\tau - i)} = \frac{\phi(i \mid H_\tau - i)}{\phi(j \mid H_\tau - i)}.$$

Enfin, si nous avons un événement de recombinaison, alors une séquence de type i sera remplacée par deux séquences de type j et k. Autrement dit, $H_{\tau+1} = H_\tau + R_i^{jk} \equiv H_\tau - i + j + k$. Nous aurons alors :

$$\begin{aligned}\frac{\phi(H_\tau)}{\phi(H_{\tau+1})} &= \frac{\phi(i \mid H_\tau - i)\phi(H_\tau - i)}{\phi(H_\tau - i + j + k)} \\ &= \frac{\phi(i \mid H_\tau - i)\phi(H_\tau - i)}{\phi(k \mid H_\tau - i + j)\phi(j|H_\tau - i)\phi(H_\tau - i)} \\ &= \frac{\phi(i \mid H_\tau - i)}{\phi(k \mid H_\tau - i + j)\phi(j \mid H_\tau - i)}.\end{aligned}$$

En résumé, nous avons :

$$\frac{\phi(H_\tau)}{\phi(H_{\tau+1})} = \begin{cases} \frac{\phi(j|H_\tau - i - j)\phi(i|H_\tau - i)}{\phi(k|H_\tau - i - j)} & \text{si} \quad H_{\tau+1} = H_\tau + C_{ij}^k, \\ \frac{\phi(i|H_\tau - i)}{\phi(j|H_\tau - i)} & \text{si} \quad H_{\tau+1} = H_\tau + M_i^j, \\ \frac{\phi(i|H_\tau - i)}{\phi(k|H_\tau - i + j)\phi(j|H_\tau - i)} & \text{si} \quad H_{\tau+1} = H_\tau + R_i^{jk}. \end{cases}$$

Il reste néanmoins à estimer la distribution $\phi(\gamma|H_\tau - \gamma)$. La façon d'approximer efficacement cette distribution ne sera pas développée ici ; le lecteur intéressé pourra se référer à Descary (2012) pour plus de détails.

4.1.4 La matrice de similarité S_{TMRCA}

Notre modèle ne permet pas d'exploiter directement la matrice $\boldsymbol{T^{hap}}$ décrite par l'équation (4.1), étant donnée que cette dernière est de taille $(2n \times 2n)$. Il est alors nécessaire de transformer cette matrice ($\boldsymbol{T^{hap}}$) en une matrice $\boldsymbol{T}$ de taille $(n \times n)$ qui donne le temps de coalescence entre individus. À notre connaissance, aucun article dans la littérature ne traite de la façon de transformer cette matrice. Alors, la façon qui nous a paru la plus simple et la plus intuitive pour le faire et que nous proposons dans ce mémoire est la suivante : le temps de coalescence entre deux individus r et s de notre échantillon est donné par le maximum des temps

de coalescence des quatre haplotypes correspondant aux individus r et s mais en excluant les temps entre les deux haplotypes du même individu. Autrement dit, pour tout $r, s \in \{1, 2, \ldots, n\}$, on définit $\boldsymbol{T}$ par ses éléments de la façon suivante :

$$T_{r,s} = \begin{cases} 0 & \text{si } \; r = s, \\ \max\{T^{hap}_{i,j}, T^{hap}_{i,j+1}, T^{hap}_{i+1,j}, T^{hap}_{i+1,j+1}\} & \text{si } \; r \neq s, \end{cases} \tag{4.12}$$

où $i = 2r - 1$ et $j = 2s - 1$.

Une fois la matrice $\boldsymbol{T}$ obtenue, il est possible de construire la matrice de similarité $\boldsymbol{S_{TMRCA}}$ comme suit :

$$S_{r,s} = \frac{\max(\boldsymbol{T}) - T_{r,s}}{\max(\boldsymbol{T})}. \tag{4.13}$$

Avant de décrire l'algorithme permettant d'obtenir la matrice $\boldsymbol{S_{TMRCA}}$, il est important de préciser que la taille d'échantillon ainsi que la longueur des séquences génétiques peuvent affecter de façon non négligeable le temps de calcul. En effet, plus les séquences sont longues, plus cela aura pour effet d'augmenter la fréquence des événements de recombinaison et qui se traduira au final par une augmentation du temps de calcul avant l'atteinte du MRCA. Pour résoudre ce problème computationnel, la solution consiste à diviser la séquence génétique en un ensemble de fenêtres composée chacune d'un nombre d de marqueurs (avec $d < p$), et de construire ensuite un ensemble d'ARG pour chacune de ces fenêtres. Ainsi, la matrice $\boldsymbol{T^{hap}}$ résultante sera donc une moyenne des moyennes des temps de coalescence. Nous pouvons à présent décrire l'algorithme qui permet d'obtenir la matrice $\boldsymbol{S_{TMRCA}}$ à partir d'un ensemble de séquences.

Algorithme TMRCA :

1. Poser H_0 l'ensemble contenant les n haplotypes de notre échantillon.
2. Pour w allant de 1 à F (le nombre de fenêtres) faire :
   (a) Pour i allant de 1 à M (le nombre de graphes que l'on veut simuler) faire :

i. Construire un ARG $G^{(i)}$ en utilisant la distribution Q de Fearnhead et Donnelly de la façon suivante :

A. Poser $\tau = 0$.

B. Tant que $\text{card}(H_\tau^{(i)}) \neq 1$ faire :

— Simuler le temps $t_\tau^{(i)}$ avant le prochain événement, avec $t_\tau^{(i)} \sim \exp\{n(n + \alpha\theta + \beta\rho - 1)/2\}$, où $n = \text{card}(H_\tau^{(i)})$.

— Calculer $Q(H_{\tau+1}^{(i)} \mid H_\tau^{(i)})$ à l'aide de l'équation (4.10) et choisir le prochain état $H_{\tau+1}^{(i)}$ proportionnellement à sa pondération.

— Poser $\tau = \tau + 1$.

C. Récupérer les temps de coalescence entre chaque paire d'haplotypes et les stocker dans une matrice $\boldsymbol{T}^{(i,w)}$.

3. Poser $\boldsymbol{T}^{hap} = (1/F \cdot M) \sum_{w=1}^{F} \sum_{i=1}^{M} \boldsymbol{T}^{(i,w)}$.

4. Créer la matrice $\boldsymbol{T}$ des temps de coalescence entre individus, tel que décrit dans l'équation (4.12).

5. Composer les éléments de la matrice $\boldsymbol{S}_{\boldsymbol{TMRCA}}$ à partir des éléments de $\boldsymbol{T}$ en utilisant la relation décrite dans l'équation (4.13).

4.2 Description du modèle et test d'association

Dans cette section, nous allons développer un nouveau test que nous avons nommé GoLiATe (pour *Genealogical and geneticaL information for Association Test*) pour tester l'association entre une région chromosomique et un phénotype. GoLiATe permet de tester l'effet combiné de plusieurs variants génétiques sur un trait quantitatif d'intérêt en utilisant à la fois l'information génétique et généalogique. L'information généalogique permet notamment de contrôler l'inflation de l'erreur de type 1 due à la structure de population et à la parenté cachée (*population structure and cryptic relatedness*).

Avant de présenter la statistique qui permet de réaliser ce test, nous allons tout

d'abord introduire quelques notations. Soit un échantillon de n séquences génétiques provenant de n individus. Notre modèle s'écrit sous la forme suivante :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{G}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{4.14}$$

où $\boldsymbol{Y}$ est un vecteur $n \times 1$ correspondant au trait quantitatif d'intérêt mesuré sur les n individus, $\boldsymbol{X} = [\mathbb{1}; \boldsymbol{X_1}; \cdots; \boldsymbol{X_q}]$ est la matrice de covariables de taille $n \times (q+1)$ et $\boldsymbol{\alpha}$ le vecteur des effets fixes correspondants. La matrice $\boldsymbol{G} = (\boldsymbol{G_1}, \ldots, \boldsymbol{G_p})$ est la matrice des génotypes constituée de n lignes correspondant aux individus et p colonnes correspondant aux marqueurs génétiques de la région d'intérêt. Les génotypes sont codés de façon à compter le nombre d'allèles mineurs pour chaque individu. Le vecteur $\boldsymbol{\beta}$ est un vecteur $p \times 1$ d'effet aléatoire qui contient les grandeurs associées à chaque variant génétique (les SNPs) et $\boldsymbol{\delta}$ un vecteur aléatoire $n \times 1$ qui permet de tenir compte de la corrélation entre les individus. On suppose que le vecteur des erreurs $\boldsymbol{\epsilon}$ et les effets aléatoires $\boldsymbol{\beta}$ et $\boldsymbol{\delta}$ sont non dépendants et distribués selon une loi normale multivariée. Autrement dit,

$$\begin{aligned}
\boldsymbol{\beta} &\sim N(0_p, \tau \boldsymbol{W}), \\
\boldsymbol{\delta} &\sim N(0_n, \tau_s \boldsymbol{S_{TMRCA}}), \\
\boldsymbol{\epsilon} &\sim N(0_n, \sigma_\epsilon^2 \boldsymbol{I_n}), \\
cov(\boldsymbol{\delta}, \boldsymbol{\epsilon}) &= 0_{n \times n}, \\
cov(\boldsymbol{\delta}, \boldsymbol{\beta}) &= 0_{n \times p}, \\
cov(\boldsymbol{\beta}, \boldsymbol{\epsilon}) &= 0_{p \times n},
\end{aligned}$$

où $\boldsymbol{W}$ est une matrice $p \times p$ diagonale contenant le poids de chaque variant génétique et $\boldsymbol{S_{TMRCA}}$ la matrice de similarité décrite précédemment. La matrice $\boldsymbol{I_n}$ correspond à la matrice identité de dimension n et $\sigma_\epsilon^2, \tau_s$ et τ sont les paramètres de la composante de variance.

Sous ces hypothèses, la variabilité phénotypique est donnée par :

$$Var(\boldsymbol{Y}) = \boldsymbol{\Omega} = \tau \boldsymbol{K} + \tau_s \boldsymbol{S_{TMRCA}} + \sigma_\epsilon^2 \boldsymbol{I_n},$$

où $\boldsymbol{K} = \boldsymbol{G}\boldsymbol{W}\boldsymbol{G}^{\boldsymbol{T}}$. La matrice $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_p)$ permet de tenir compte de la contribution de chaque variant génétique de la région sous étude. Pour les variants rares, on détermine les poids en utilisant la densité de la loi bêta de paramètres $a_1 = 1$ et $a_2 = 25$ évaluée à la fréquence d'allèle mineur (MAF) de chaque variant (Wu *et al.*, 2011). Autrement dit, $\sqrt{w_j} = \text{Beta}\,(MAF_j; 1, 25)$. Pour des analyses avec des variants communs ($MAF \geq 5\%$), on peut utiliser un poids unitaire pour tous les variants ($\boldsymbol{W} = \boldsymbol{I_p}$) et qui correspond finalement à $\sqrt{w_j} = \text{Beta}\,(MAF_j; 1, 1)$.

Pour tester l'effet de la région $(\boldsymbol{G_1}, \ldots, \boldsymbol{G_p})$ sur le phénotype $\boldsymbol{Y}$, nous suggérons le test d'hypothèse

$$\text{H}_0 : \tau = 0 \quad \text{contre} \quad \text{H}_a : \tau > 0.$$

Pour ce faire, nous utilisons la statistique de score. L'avantage d'utiliser un test de score, comme nous l'avons souligné auparavant, est qu'il nécessite seulement d'ajuster le modèle sous l'hypothèse nulle.
Soit $\boldsymbol{\theta} = (\boldsymbol{\alpha}^{\boldsymbol{T}}, \tau_s, \sigma_\epsilon^2)^T$, la fonction log-vraisemblance du modèle de l'équation (4.14) s'écrit comme suit :

$$l(\boldsymbol{\theta}, \tau; \boldsymbol{Y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})^{\boldsymbol{T}}\boldsymbol{\Omega}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}).$$

Pour avoir la statistique score du test $H_0 : \tau = 0$, on commence tout d'abord par trouver la dérivée de $l(\boldsymbol{\theta}, \tau; \boldsymbol{Y})$ par rapport à τ. Celle ci est donnée par :

$$\frac{\partial l(\boldsymbol{\theta}, \tau; \boldsymbol{Y})}{\partial \tau} = -\frac{1}{2}tr(\boldsymbol{\Omega}^{-1}\boldsymbol{K}) + \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})^{\boldsymbol{T}}\boldsymbol{\Omega}^{-1}\boldsymbol{K}\boldsymbol{\Omega}^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}). \quad (4.15)$$

Puisque le premier terme de l'équation (4.15) ne dépend pas des données (le phénotype d'intérêt $\boldsymbol{Y}$) et que la matrice $\boldsymbol{\Omega}$ est estimée de façon robuste alors, ce terme ne sera pas considéré dans la statistique du test. Ainsi, comme dans le modèle SKAT (Wu et al. ; 2011), la statistique que nous utilisons correspond à

deux fois le second terme de l'équation (4.15) et est donnée par :

$$Q_\tau = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})^T \boldsymbol{\Omega}^{-1} \boldsymbol{K} \boldsymbol{\Omega}^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}) \Big|_{\tau=0, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{4.16}$$

où $\hat{\boldsymbol{\theta}}$ est l'estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ du modèle (4.14) sous l'hypothèse nulle, autrement dit, le modèle :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\delta} + \boldsymbol{\epsilon}. \tag{4.17}$$

Afin d'estimer $\boldsymbol{\theta}$, nous devons écrire la fonction log-vraisemblance sous H_0, soit sous le modèle (4.17). Celle-ci est donnée par :

$$l_0(\boldsymbol{\theta}; \boldsymbol{Y}) = -\frac{n}{2} \log(2\pi\tau_s) - \frac{1}{2} \log |\boldsymbol{\Omega}_\eta| - \frac{1}{2\tau_s} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha})^T \boldsymbol{\Omega}_\eta^{-1} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}), \tag{4.18}$$

où

$$\boldsymbol{\Omega}_\eta = \boldsymbol{S}_{\boldsymbol{TMRCA}} + \frac{\sigma_\epsilon^2}{\tau_s} \boldsymbol{I}_n = \boldsymbol{S}_{\boldsymbol{TMRCA}} + \eta \boldsymbol{I}_n.$$

En annulant la première dérivée de $l_0(\boldsymbol{\theta}; \boldsymbol{Y})$ par rapport à $\boldsymbol{\alpha}$, nous obtenons :

$$\boldsymbol{\alpha}(\eta) = (\boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{Y}. \tag{4.19}$$

Pour touver la valeur du maximum de vraisemblance de τ_s comme fonction de η, la valeur du maximum de vraisemblance des effets fixes $\boldsymbol{\alpha}(\eta)$ qui ne dépend pas de τ_s est remplacée dans (4.18), puis en annulant la première dérivée par rapport à τ_s, nous obtenons :

$$\begin{aligned}
\tau_s(\eta) &= \frac{1}{n} \{\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}(\eta)\}^T \boldsymbol{\Omega}_\eta^{-1} \{\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\alpha}(\eta)\} \\
&= \frac{1}{n} \left\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{Y}\right\}^T \boldsymbol{\Omega}_\eta^{-1} \\
&\quad \times \left\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{Y}\right\} \\
&= \frac{1}{n} \left[\left\{\boldsymbol{I}_n - X(\boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1}\right\} \boldsymbol{Y}\right]^T \boldsymbol{\Omega}_\eta^{-1} \\
&\quad \times \left[\left\{\boldsymbol{I}_n - X(\boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{\Omega}_\eta^{-1}\right\} \boldsymbol{Y}\right] \\
&= \frac{1}{n} \boldsymbol{Y}^T \boldsymbol{H}_\eta^T \boldsymbol{\Omega}_\eta^{-1} \boldsymbol{H}_\eta \boldsymbol{Y},
\end{aligned} \tag{4.20}$$

où $\boldsymbol{H}_\eta = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{\Omega}_\eta^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{\Omega}_\eta^{-1}$. Encore une fois, en substituant (4.20) et (4.19) dans (4.18), nous obtenons la fonction log-vraisemblance profilée :

$$l_0(\eta; \boldsymbol{Y}) = -\frac{n}{2}\left(1 + \log\frac{2\pi}{n}\right) - \frac{1}{2}\log|\boldsymbol{\Omega}_\eta| - \frac{n}{2}\log \boldsymbol{Y}^T\boldsymbol{H}_\eta^T\boldsymbol{\Omega}_\eta^{-1}\boldsymbol{H}_\eta\boldsymbol{Y}, \qquad (4.21)$$

qui ne dépend plus que d'un seul paramètre η. Il ne reste plus qu'à optimiser (4.21) par rapport à η en utilisant un algorithme numérique. Il suffit ensuite de remplacer la valeur de η dans (4.19) et (4.20) pour obtenir $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\alpha}}^T, \hat{\tau}_s, \hat{\sigma}_\epsilon^2)^T$. Ainsi, nous estimons $\boldsymbol{\Omega}$ et $\boldsymbol{\alpha}$ par :

$$\begin{aligned}
\widehat{\boldsymbol{\Omega}}_0 &= \hat{\tau}_s\boldsymbol{S}_{TMTCA} + \hat{\sigma}_\epsilon^2\boldsymbol{I}_n, \\
\hat{\boldsymbol{\alpha}} &= (\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{Y}.
\end{aligned}$$

À présent, pour trouver la distribution de Q_τ sous l'hypothèse nulle, nous allons réécrire Q_τ comme suit :

$$\begin{aligned}
Q_\tau &= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}})^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Omega}}_0^{-1}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}) \\
&= \left\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{Y}\right\}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Omega}}_0^{-1} \\
&\quad \times\left\{\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{Y}\right\} \\
&= \left[\left\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\right\}\boldsymbol{Y}\right]^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{K}\widehat{\boldsymbol{\Omega}}_0^{-1} \\
&\quad \times\left[\left\{\boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\right\}\boldsymbol{Y}\right] \\
&= \boldsymbol{Y}^T\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\boldsymbol{Y} \\
&= \boldsymbol{Y}^T\widehat{\boldsymbol{\Omega}}_0^{-1/2}\widehat{\boldsymbol{\Omega}}_0^{1/2}\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\widehat{\boldsymbol{\Omega}}_0^{1/2}\widehat{\boldsymbol{\Omega}}_0^{-1/2}\boldsymbol{Y} \\
&= \widetilde{\boldsymbol{Y}}^T\widehat{\boldsymbol{\Omega}}_0^{1/2}\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\widehat{\boldsymbol{\Omega}}_0^{1/2}\widetilde{\boldsymbol{Y}}, \qquad (4.22)
\end{aligned}$$

où $\boldsymbol{P} = \widehat{\boldsymbol{\Omega}}_0^{-1} - \widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{X}(\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{\Omega}}_0^{-1}$ est une matrice symétrique idempotente et $\widetilde{\boldsymbol{Y}} = \widehat{\boldsymbol{\Omega}}_0^{-1/2}\boldsymbol{Y} \sim N(0_n, \boldsymbol{I}_n)$.

Soit $\boldsymbol{\Lambda} = \text{diag}\,(\lambda_1, \ldots, \lambda_p)$ tel que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$, la matrice diagonale des valeurs propres associées à la matrice $\widehat{\boldsymbol{\Omega}}_0^{1/2}\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\widehat{\boldsymbol{\Omega}}_0^{1/2}$ et $\boldsymbol{U}$ la matrice des

vecteurs propres associés. Ainsi, nous avons :

$$\widehat{\boldsymbol{\Omega}}_0^{1/2}\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\widehat{\boldsymbol{\Omega}}_0^{1/2} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T \tag{4.23}$$

En insérant le résultat de l'équation (4.23) dans l'équation (4.22), nous obtenons :

$$\widetilde{\boldsymbol{Y}}^T\widehat{\boldsymbol{\Omega}}_0^{1/2}\boldsymbol{P}\boldsymbol{K}\boldsymbol{P}\widehat{\boldsymbol{\Omega}}_0^{1/2}\widetilde{\boldsymbol{Y}} = \widetilde{\boldsymbol{Y}}^T\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T\widetilde{\boldsymbol{Y}} = \sum_{i=1}^{p}\lambda_i Z_i^2, \tag{4.24}$$

où $\boldsymbol{Z} = \boldsymbol{U}^T\widetilde{\boldsymbol{Y}} = (Z_1, \cdots, Z_p)^T \sim N(0_p, \boldsymbol{I}_p)$ et $Z_i^2 \sim \chi_1^2$. Ainsi, la loi de Q_τ est un mélange de lois de khi-deux à 1 degré de liberté.

Pour obtenir la valeur-p associée à Q_τ, on utilise l'approximation de Davies (Davies, 1980). Cette méthode approxime le quantile d'une forme quadratique $Q = \boldsymbol{Z}^T\boldsymbol{\Lambda}\boldsymbol{Z}$ (où $\boldsymbol{Z}$ est un vecteur normal) par une intégrale calculée numériquement. Cette approximation de Davies est basée sur l'inversion de la fonction caractéristique de l'équation (4.24).

CHAPITRE V

SIMULATIONS ET RÉSULTATS

Afin d'étudier la performance de notre modèle GoLiATe, nous avons procédé à une analyse par simulation. Cela nous a permis de tester plusieurs scénarios mais aussi de nous assurer que nos données respectent bien certaines hypothèses notamment la présence d'une structure de population de notre échantillon. Ainsi, nous avons pu comparer les performances de GoLiATe en termes de puissance et de protection contre l'erreur de type 1 à celles de SKAT (Wu *et al.*, 2011), de SKAT-O (Lee *et al.*, 2012) et de MiST (Sun *et al.*, 2013), que nous avons introduit à la section 3.5 du chapitre III.

Avant de présenter les résultats, nous allons tout d'abord commencer par décrire la façon dont nous avons simulé nos données.

5.1 Simulation et préparation des données

Pour simuler un échantillon de séquences génétiques, nous avons eu recours à l'utilisation d'un programme nommé HAPSIMU (Zhang *et al.*, 2008). Ce programme utilise une partie de la base de données réelles ENCODE du projet HapMap pour simuler des séquences génétiques (les génotypes des individus). Cette base de données est constituée d'un ensemble de SNPs tirés de dix régions autosomales différentes de 500 Kb chacune et provenant de quatre populations différentes.

Le programme HAPSIMU utilise des haplotypes réels de la base de données ENCODE pour simuler des populations hétérogènes avec diverses structures connues et contrôlables. Cette base de données est contituée d'individus Africains YRI, (de l'anglais *Yoruba from Ibadan of Africa*), et d'individus de type caucasien avec des ancêtres nord Européens et de l'Europe de l'ouest CEPH, (de l'anglais *Caucasian with northern and western European ancestry*). Au total, 12867 SNPs ont été sélectionnés sur les dix régions. Pour simuler un échantillon de séquences génétiques, le programme HAPSIMU simule dans un premier temps, 1000 individus CEPH et 1000 individus YRI à partir des haplotypes ENCODE pour être utilisés comme populations fondatrices. Par la suite, la population hétérogène composée de CEPH et de YRI est simulée selon un modèle de migration discret pour le mélange des populations. Autrement dit, les 1000 individus fondateurs de chacune des sous-populations (CEPH et YRI) sont accouplés de façon aléatoire à l'intérieur de leurs populations respectives jusqu'à ce que le nombre de générations de descendants voulu soit atteint. Le fait de simuler plusieurs générations contribue à créer davantage de diversité génétique chez les individus. La valeur par défaut du nombre de générations est de 5 (valeur que nous avons utilisé pour notre simulation). Au cours de ce processus, la taille de la population est maintenue constante à 1000 pour chaque génération et tous les marqueurs sont supposés être en équilibre de Hardy-Weinberg et sont recombinés de façon aléatoire selon les fractions de recombinaison estimées à l'aide de la fonction de Kosambi (1944) à partir des données réelles. Une fois les deux sous-populations de CEPH et de YRI obtenues, il est alors possible de former un échantillon de séquences composé d'une proportion p_{yri} de YRI et $1 - p_{yri}$ de CEPH spécifiée à l'avance.

Pour tester notre modèle, nous avons choisi de simuler un échantillon de 200 individus (100 YRI et 100 CEPH) correspondant à 200 séquences génétiques composées de 12867 marqueurs représentant 10 régions autosomales de 500 Kb chacune. Puisque ces séquences contiennent un très grand nombre de marqueurs et qu'il

n'est pas possible de construire des ARGs avec tous ces marqueurs en un temps raisonnable, nous avons sélectionné seulement $p = 100$ marqueurs correspondant à une région de 41 Kb provenant de la région 7q21.13 [1] du chromosome 7. Le choix de cette région n'a pas été fait de façon aléatoire, mais en nous assurons qu'elle contienne bien des marqueurs informatifs pour la structure de population. Ces marqueurs (informatifs) sont ceux dont la fréquence de l'allèle mineur est significativement différente entre les deux sous-populations de CEPH et de YRI dans l'échantillon. Comme le programme HAPSIMU permet d'identifier l'appartenance de chaque individu de l'échantillon à une sous-population (CEPH ou YRI), nous avons donc pu identifier les marqueurs informatifs à l'aide d'un test d'égalité des proportions. La figure 5.1 présente les résultats du test d'égalité des proportions, ainsi que la distribution des marqueurs informatifs identifiés pour chacune des sous-populations à l'aide d'une boîte à moustaches (*box plot*). Le graphique de gauche montre la valeur du $-\log_{10}$(valeur-p) du test pour chaque marqueur de la région 7q21.13, ainsi que le seuil (représenté par une ligne horizontale de couleur rouge), au delà duquel l'hypothèse d'égalité est rejetée. Il est à noter que ce seuil a été calculé à l'aide de la correction de Bonferroni pour les tests multiples. Le graphique de droite illustre la distribution des fréquences d'allèles mineurs des marqueurs significativement différents identifiés sur le graphique de gauche.
En analysant ces *box plot*, nous constatons que ces marqueurs distinguent bien les deux sous-populations de CEPH et de YRI présentes dans notre échantillon. De plus, nous remarquons que certains variants génétiques ont une fréquence d'allèle mineur supérieure à 0.5. Cela est principalement dû au fait qu'en stratifiant l'échantillon, certains marqueurs apparaissent seulement chez les individus d'une sous-population et pas dans l'autre.

---

1. Le 7 indique la $7^{ème}$ paire de chromosomes ; le q indique qu'il se situe sur le bras long du chromosome et le 21.13 désigne la position exacte de la région par rapport au centromère.

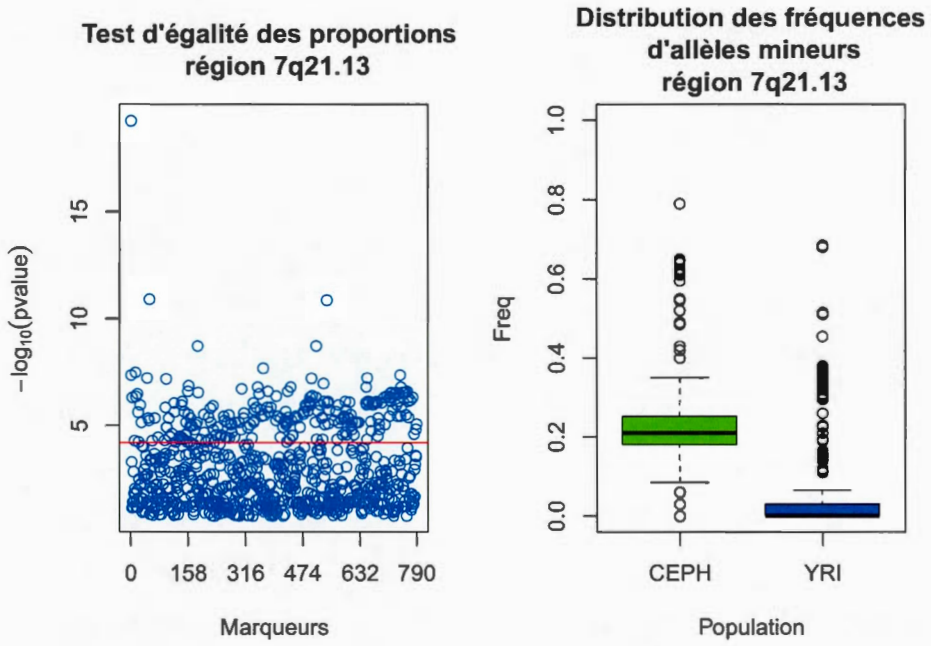


Figure 5.1 Représentation des résultats du test d'égalité des proportions et de la distribution des fréquences d'allèles mineurs des marqueurs informatifs pour la structure de population pour la région 7q21.13 du chromosome 7.

Les données simulées de HAPSIMU sont contenues dans deux fichiers distincts avec des extensions ".ped" et ".map" lisible par le programme PLINK [2]. Le fichier map (*map file*) contient le nom de la région, les noms des marqueurs ainsi que les distances génétiques entre chaque marqueur. Quant au fichier ped (*ped file*), il contient la matrice des génotypes correspondant aux 200 individus de notre échantillon. Comme le programme HAPSIMU ne permet pas d'obtenir les haplotypes des individus de notre échantillon nécessaire à l'obtention de la matrice

2. PLINK est un programme disponible gratuitement conçu pour effectuer de nombreuses analyses statistiques sur le génome en entier de façon très efficace. Plus d'informations sont disponibles sur le site `http://pngu.mgh.harvard.edu/~purcell/plink/`

S_{TMRCA}, il nous a donc fallu utiliser un autre programme qui permet de reconstruire les haplotypes à partir des génotypes observés des individus de l'échantillon. Cette étape a été effectuée grâce au programme IMPUTE2 qui utilise des algorithmes MCMC (*Markov Chain Monte Carlo*) pour obtenir les haplotypes les plus probables pour chaque individu. Cette étape nécessite l'utilisation de cartes chromosomiques provenant de la phase 3 du projet "1000 génomes", et servant de référence dans la reconstitution des haplotypes. Les développements théoriques relatifs au programme IMPUTE2 sont complexes et ne seront donc pas exposés dans ce mémoire. Le lecteur intéressé peut se référer à l'article de Howie *et al.* (2009) pour plus de détails. Une fois les haplotypes obtenus, nous avons utilisé le programme TMRCA dont l'algorithme a été décrit à la section 4.1.4 du chapitre précédent pour obtenir la matrice de similarité S_{TMRCA}. Nous disposions donc d'une région de 41 Kb constituée de 100 marqueurs que nous avons divisé en 18 fenêtres de 10 marqueurs chacune. Le nombre de graphes simulés pour chaque fenêtre est de $M = 1000$ graphes. Ainsi, 18000 graphes ont été simulés au total pour obtenir la matrice S_{TMRCA}.

Pour simuler le phénotype Y_i de l'individu i, nous avons utilisé le modèle additif suivant :

$$Y_i = \mu + \sum_{j=1}^{k} \beta_j G_{ij}^c + \epsilon_i, \tag{5.1}$$

où μ est une constante, $\epsilon_i \sim N(0,1)$ et $(G_{i1}^c, \ldots, G_{ik}^c)$ sont les génotypes des k variants de la région associés au phénotype et β_j les coefficients correspondants tels que :

$$\beta_j = \sqrt{\frac{v_j}{2MAF_j(1-MAF_j)}},$$

où v_j représente la variabilité phénotypique expliquée par le $j^{\text{ème}}$ variant génétique associé et MAF_j sa fréquence d'allèle mineur.

5.2 Évaluation du modèle

5.2.1 Évaluation de l'erreur de type 1

Pour tester si notre modèle est valide en terme d'erreur de type 1, nous avons simulé des données sous l'hypothèse nulle de non association entre la région d'intérêt 7q21.13 et le phénotype. Un ensemble de marqueurs provenant de la région 12q12 du chromosome 12 informative pour la structure de population a d'abord servi à simuler le phénotype $\boldsymbol{Y}$ selon le modèle additif décrit par l'équation (5.1) sous l'hypothèse H_0 :"$\boldsymbol{Y}$ n'est pas associé à la région 7q21.13". Au total, trois marqueurs de la région 12q12 informatifs pour la structure de population ont été associés au phénotype en fixant $\mu = 3$, $\beta_1 = 0.0812$, $\beta_2 = 0.1429$ et $\beta_3 = 0.0858$. Par la suite, le phénotype obtenu est testé avec les marqueurs sélectionnés provenant de la région 7q21.13 citée précédemment. L'avantage de procéder ainsi est que d'une part nous nous assurons qu'il n'y a aucune association entre notre région d'intérêt (région 7q21.13) et le phénotype simulé et, d'autre part, cela permet de conserver l'information de la structure présente dans les données. Cette méthodologie est inspirée de celle utilisée par Liu *et al.* (2013) appliquée pour des plans d'études cas-témoins.

Dix mille simulations ont été réalisées pour évaluer l'erreur de type 1. Cette dernière est calculée de façon empirique par la proportion de valeurs-p ayant une valeur inférieure au seuil α considéré. Ainsi, nous avons pu comparer notre modèle GoLiATe avec MiST, SKAT-O et SKAT en utilisant différents noyaux pour SKAT, notamment notre matrice de similarité $\boldsymbol{S_{TMRCA}}$, ainsi que le noyau IBS (*identity by state*). Ces derniers sont notés par SKAT_$\boldsymbol{S_{TMRCA}}$ et SKAT_IBS respectivement. Le noyau IBS permet de tenir compte de la similarité entre individus en utilisant l'information du nombre d'allèles identiques par états qu'ils partagent entre eux. Le noyau linéaire pondéré a été utilisé pour MiST et SKAT-O en fixant les poids des variants à l'aide de la distribution béta ($\sqrt{w_j} = \text{Beta}(MAF_j, 1, 25)$).

Pour GoLiATe, nous avons fixé les poids de la façon suivante

$$\sqrt{w_j} = \begin{cases} \text{Beta}(MAF_j; 1, 25) & \text{si} \quad MAF_j \leq 0.08, \\ \sqrt{13 - (24 \times MAF_j)} & \text{si} \quad MAF_j > 0.08. \end{cases}$$

Cette façon de faire nous permet d'accorder plus de poids aux variants rares mais sans négliger la présence des variants communs. En effet, comme la loi bêta accorde un poids pratiquement nul aux variants dont la fréquence d'allèles est supérieur à 0.2, nous avons plutôt choisi une fonction qui décroît graduellement vers 1, afin que ces variants puissent contribuer à la statistique du test.

Les résultats des simulations pour l'évaluation de l'erreur de type 1 sont présentés dans le tableau 5.1. Les "quantile-quantile plots" des valeurs-p observées en fonction des valeurs-p espérées (uniformément distribuées sous l'hypothèse nulle) pour les différentes méthodes, sont illustrés à la figure 5.1. Ces résultats montrent que les valeurs-p de GoLiATe et SKAT_$\boldsymbol{S_{TMRCA}}$ suivent bien la distribution uniforme (bon alignement sur l'axe $X = Y$) sous l'hypothèse nulle à l'opposé de SKAT-O, SKAT_IBS et MiST. Nous pouvons donc dire que GoLiATe est un test valide en terme d'erreur de type 1 et peut donc être utilisé sans avoir besoin de recourir aux méthodes de correction de l'effet de structure de population.

Tableau 5.1 Résultats d'estimation de l'erreur de type 1 sur la région 7q21.13. Le tableau montre la proportion de valeurs-p inférieure au seuil α en utilisant 10,000 simulations. Notre modèle GoLiATe a été comparé avec SKAT_$\boldsymbol{S_{TMRCA}}$, SKAT_IBS, SKAT-O et MiST.

seuil α	GoLiATe	SKAT_$\boldsymbol{S_{TMRCA}}$	SKAT_IBS	SKAT-O	MiST
0.01	0.0087	0.0117	0.0196	0.0169	0.0193
0.025	0.0232	0.0262	0.0459	0.0353	0.0416
0.05	0.0475	0.0517	0.0850	0.0658	0.0820
0.1	0.1004	0.1046	0.1600	0.1250	0.1479

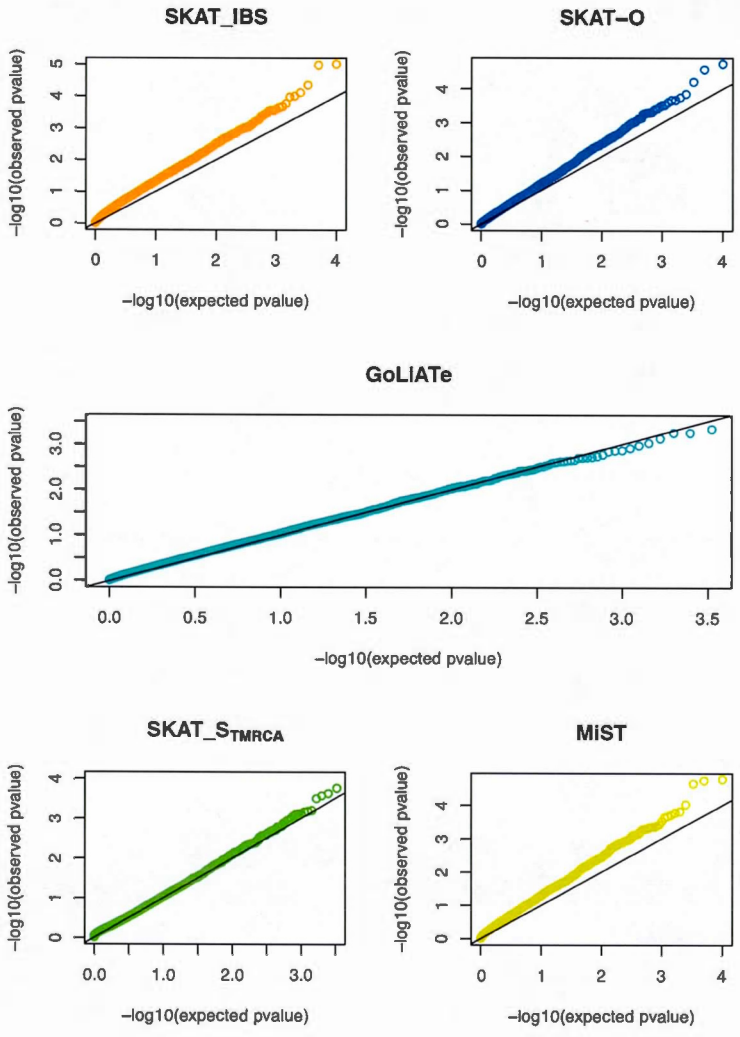


Figure 5.2 Quantile-Quantile plots de la distribution des p-values sous l'hypothèse nulle de la région 7q21.13 du chromosome 7 pour les modèles GoLiATe, SKAT_IBS, SKAT_S_{TMRCA}, SKAT-O et MiST obtenues à partir de 10,000 simulations. $-\log_{10}$ des valeurs-p observées sont représentées en fonction de leurs valeurs espérées.

5.2.2 Évaluation de la puissance

Avant de présenter les différents scénarios pour évaluer et comparer la puissance de GoLiATe avec la puissance des tests SKAT, SKAT-O et MiST, il est important de souligner que ces derniers ont montré une inflation de l'erreur de type 1 à cause de la présence d'une structure de population dans notre échantillon. Ainsi, pour rendre les résultats de GoLiATe comparables avec ceux des autres tests, nous avons dû ajuster ces derniers en incluant comme covariables les composantes principales de la matrice *kinship* empirique $\boldsymbol{\Phi}$, décrite à la section 3.4.2 du chapitre III, afin de contrôler l'effet de la structure de population. On se souvient que lorsque l'on avait simulé nos séquences génétiques, nous disposions de dix régions autosomales de 500Kb chacune avec un nombre total de 12867 marqueurs. À partir de ces marqueurs, nous avons sélectionné 3748 SNPs informatifs pour la structure de population identifiés à l'aide d'un test d'égalité des proportions tel qu'il a été décrit dans la section 5.1 du présent chapitre, afin de construire la matrice $\boldsymbol{\Phi}$. Comme notre échantillon est formé d'individus provenant de deux sous-populations seulement, alors, seule la première composante principale correspondant au vecteur propre associé à la plus grande valeur propre de la matrice $\boldsymbol{\Phi}$ suffit pour contrôler l'effet de la stratification.

Pour vérifier que l'on parvient effectivement à contrôler l'effet de la stratification avec les composantes principales, nous avons de nouveau simulé des données sous l'hypothèse nulle de non association de la région 7q21.13 avec le phénotype pour les modèles SKAT-O, SKAT_IBS et MiST. Les résultats sont présentés à la figure 5.3. En analysant les "quantile-quantile plots", nous constatons que les valeurs-p observées des tests SKAT-O, SKAT_IBS et MiST suivent bien la distribution uniforme (bon alignement sur l'axe $X = Y$) sous l'hypothèse nulle après la correction par les composantes principales.

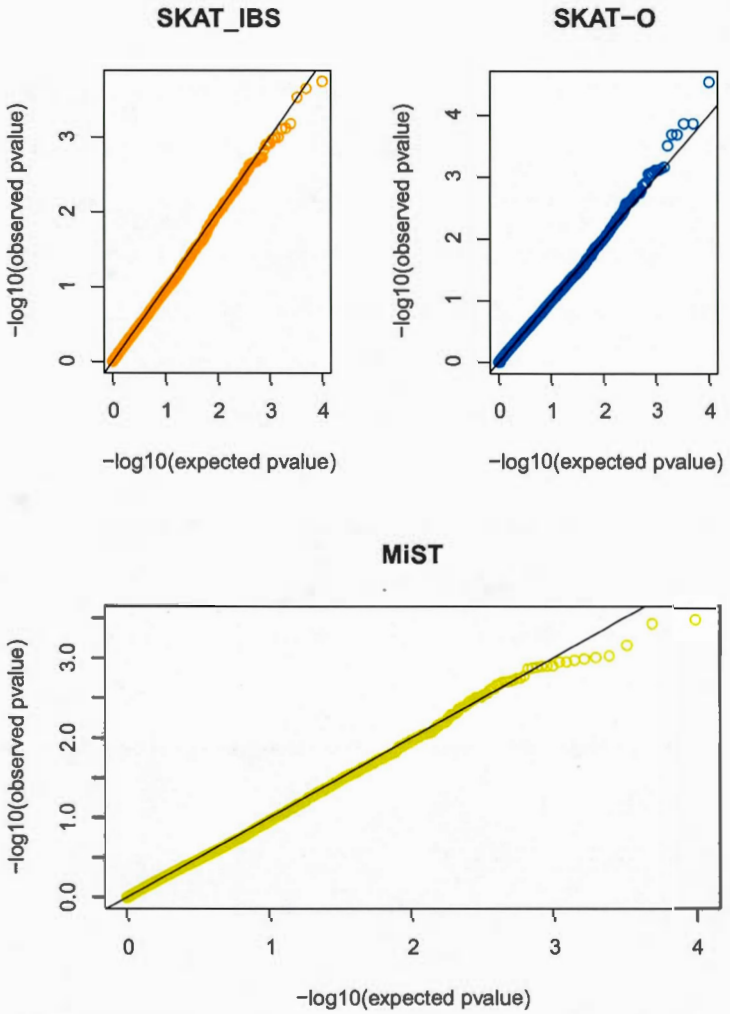


Figure 5.3 Quantile-Quantile plots de la distribution des p-values sous l'hypothèse de non association avec la région 7q21.13 du chromosome 7, après l'ajustement par les composantes principales, pour les modèles SKAT_IBS, SKAT-O et MiST obtenues à partir de 10,000 simulations. $-\log_{10}$ des valeurs-p observées sont représentées en fonction de leurs valeurs espérées.

À présent, nous sommes en mesure de procéder à l'évaluation de la puissance des tests. À cette fin, nous avons simulé des données sous l'hypothèse alternative d'association entre la région d'intérêt 7q21.13 et le phénotype. Ce dernier a été simulé selon le modèle additif décrit par l'équation (5.1) en associant un ensemble de marqueurs provenant de la région 7q21.13. Ainsi, nous avons pu tester plusieurs scénarios. Ces derniers sont résumés dans le tableau 5.2.

Tableau 5.2 Tableau résumant les caractéristiques de chaque scénario de simulation pour l'évaluation de la puissance. Le nombre d'individus de l'échantillon est représenté par n, le nombre de variants dans la région par p et la proportion de variants causaux par p_{causal}. Enfin, h^2 représente la variabilité phénotypique totale expliquée par la région.

Région 7q21.13 : 41Kb				
Scénario	$(n;p)$	p_{causal}	type de variants associés	h^2
1	(200 ;100)	3%	100% variants communs (VC)	1%
2	(200 ;100)	3%	100% variants rares (VR)	1%
3	(200 ;100)	3%	65% VC- 35% VR	1%
4	(200 ;100)	5%	100% variants communs (VC)	1%
5	(200 ;100)	5%	100% variants rares (VR)	1%
6	(200 ;100)	5%	60% VC- 40% VR	1%
7	(200 ;100)	10%	100% variants communs (VC)	1%
8	(200 ;100)	10%	100% variants rares (VR)	1%
9	(200 ;100)	10%	60% VC- 40% VR	1%
10	(200 ;100)	3%	100% variants communs (VC)	1.5%
11	(200 ;100)	3%	100% variants rares (VR)	1.5%
12	(200 ;100)	3%	65% VC- 35% VR	1.5%
13	(200 ;100)	5%	100% variants communs (VC)	1.5%
14	(200 ;100)	5%	100% variants rares (VR)	1.5%
15	(200 ;100)	5%	60% VC- 40% VR	1.5%
16	(200 ;100)	10%	100% variants communs (VC)	1.5%
17	(200 ;100)	10%	100% variants rares (VR)	1.5%
18	(200 ;100)	10%	60% VC- 40% VR	1.5%

Rappelons que nous avons sélectionné seulement une centaine de marqueurs correspondant à une taille de région de 41 Kb, constituée à 44% de variants rares, dont la fréquence d'allèles est inférieure à 5%. Comme pour l'évaluation de l'erreur de type 1, nous avons utilisé le noyau IBS et notre matrice de similarité $\boldsymbol{S_{TMRCA}}$ pour SKAT et le noyau linéaire pondéré pour GoLiATe, SKAT-O et MiST. Le poids de chaque variant a été déterminé de la même façon présentée à la section précédente pour l'évaluation de l'erreur de type 1.

Ainsi, en effectuant 10,000 simulations, nous avons évalué la puissance de chaque test par la proportion de valeur-p dont la valeur est inférieure au seuil $\alpha = 0.05$. Les résultats pour les différents scénarios décrits dans le tableau 5.2 sont présentés dans la figure 5.4. Cette figure est constituée de six graphiques. Chacun de ces graphiques présente simultanément les résultats sur la puissance pour trois scénarios différents en variant la proportion de marqueurs causaux. Les trois graphiques de gauche montrent les résultats pour les scénarios 1 à 9 et ceux de droite pour les scénarios 10 à 18. Ces résultats montrent clairement un manque de puissance pour l'ensemble des méthodes considérées en raison de taille d'échantillon limitée à 200 individus. Toutefois, en observant les résultats des scénarios 2, 5, 8, 11, 14 et 17 présentés dans le tableau 5.2, c'est à dire les scénarios où nous avons associé uniquement des variants rares (voir les deux graphiques en haut de la figure 5.4), nous constatons que GoLiATe semble avoir une puissance comparable à celle des autres tests. La matrice $\boldsymbol{S_{TMRCA}}$ utilisée seule comme noyau dans SKAT ne donne pas de bons résultats. En effet, dans la majorité des cas, SKAT_$\boldsymbol{S_{TMRCA}}$ montre une puissance inférieure à celle des autres tests sauf pour les scénarios 3, 12 et 18 où la proportion de variants associés est faible (3%).

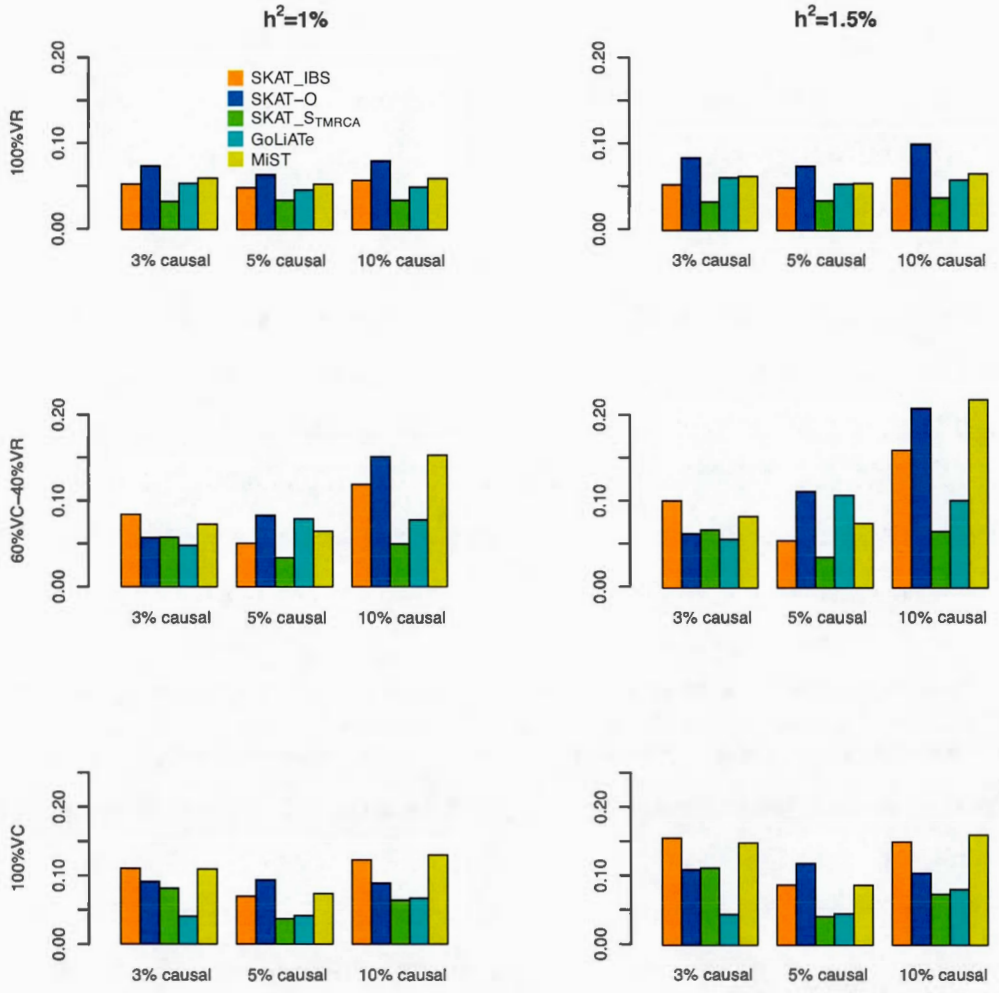


Figure 5.4 Figure illustrant les résultats de la puissance des tests GoLiATe, SKAT_S_{TMRCA}, SKAT_IBS, SKAT-O et MiST, obtenus à partir de 10,000 simulations pour les scénarios 1 à 18, sur la région 7q21.13 du chromosome 7. Pour chaque méthode la puissance est évaluée par la proportion de valeur-p inférieure au seuil $\alpha = 0.05$.

CONCLUSION

Ce mémoire avait pour objectif de développer un nouveau test d'association en combinant les modèles linéaires mixtes et le processus de coalescence. Ce dernier nous a permis de développer une nouvelle matrice de similarité basée sur le temps à l'ancêtre commun, que nous avons nommée S_{TMRCA}, et qui permet de tenir compte de la dépendance ancestrale entre les individus de l'échantillon. Cette matrice a été par la suite introduite dans la structure de variance d'un modèle linéaire mixte afin d'essayer de contrôler l'inflation de l'erreur de type 1 due à la structure de population, mais aussi pour éventuellement contribuer à la puissance du test en apportant de l'information généalogique additionnelle.

Nous avons donc développé la vraisemblance de notre modèle et présenté un test de score. Ce dernier nécessitait l'estimation de paramètres sous l'hypothèse nulle de non association qui ont été évalués de façon robuste en utilisant une vraisemblance profilée. Enfin, nous avons simulé des données sous l'hypothèse nulle et alternative, afin d'évaluer et de comparer l'erreur de type 1 et la puissance de notre test avec celles des tests présentés à la section 3.5 du chapitre III.

Les résultats présentés à la section 5.2.1 suggèrent que notre test GoLiATe a un bon contrôle de l'erreur de type 1 en présence d'une structure de population (due à la stratification) dans l'échantillon, à l'opposé des tests SKAT, SKAT-O et MiST. Ainsi, nous pouvons dire que notre matrice de similarité S_{TMRCA} est informative pour la structure de population. De plus, les résultats présentés à la section 5.2.2 montrent que GoLiATe semble être plus adapté à un contexte de variants rares qu'à un contexte de variants communs. En effet, dans la majorité des cas où nous

avons associé des variants rares, GoLiATe montre une puissance comparable avec celle de SKAT et de MiST.

Si nous avions eu plus de temps, il aurait été par exemple intéressant de comparer les performances de GoLiATe dans un scénario où la structure de population est due à la *cryptic relatedness* (CR) et non à la stratification. En effet, dans ce cas, la correction de l'effet de la structure avec les composantes principales pour les tests SKAT, SKAT-O et MiST ne fonctionne pas comme dans le cas de la stratification (Astle et Balding, 2009). Nous pensons donc que le gain de GoLiATe par rapport aux autres méthodes serait plus important dans un contexte de CR que dans un contexte de stratification. De plus, il serait intéressant de construire un test qui permet de tester simultanément si les composantes de variance du modèle (4.13) sont significativement différentes de zéro. Autrement dit, effectuer le test $H_0 : \tau = 0, \tau_s = 0$ au lieu de $H_0 : \tau = 0$ à l'aide d'un test de score bivarié par exemple ; le défi étant de trouver la distribution du vecteur-score bivarié sous l'hypothèse nulle. En effet, en utilisant notre matrice de similarité $\boldsymbol{S_{TMRCA}}$ comme noyau dans SKAT, nous avons tout de même pu détecter un signal sous l'hypothèse d'association. En d'autres termes, en plus de la capacité dont dispose notre matrice $\boldsymbol{S_{TMRCA}}$ à capter l'information de la structure de population, elle peut aussi servir à la détection de la véritable association entre le caractère étudié et la région chromosomique d'intérêt ; en gardant à l'esprit que l'information du temps à l'ancêtre commun qu'utilise la matrice $\boldsymbol{S_{TMRCA}}$ est complètement différente de celle contenue dans les génotypes.

Il reste cependant beaucoup à faire afin d'améliorer les performances de GoLiATe, et de combiner de la meilleure façon l'information ancestrale contenue dans $\boldsymbol{S_{TMRCA}}$ avec celle des génotypes.

RÉFÉRENCES

Astle, W. et Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, 451–471.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, *7*(10), 781–791.

Benjamin A. Pierce, R. C. (2012). *L'essentiel de la génétique.* De Boeck.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P. et Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature protocols*, *6*(2), 121–133.

Davies, R. B. (1980). Algorithm as 155 : The distribution of a linear combination of χ 2 random variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *29*(3), 323–333.

Descary, M.-H. (2012). *Dmap : Une nouvelle méthode de cartographie génétique fine adaptée à des modèles génétiques complexes.* (Mémoire de maîtrise). Université du Québec À Montréal.

Dupont, M. (2013). *Cartographie génétique fine : évaluation d'une méthode d'estimation des allèles et du modèle de pénétrance.* (Mémoire de maîtrise). Université du Québec À Montréal.

Fearnhead, P. et Donnelly, P. (2001). Estimating recombination rates from population genetic data. *Genetics*, *159*(3), 1299–1318.

Forest, M. (2010). *Cartographie génétique fine simultanée de deux gènes.* (Mémoire de maîtrise). Université du Québec À Montréal.

Griffiths, R. C. et Marjoram, P. (1996). Ancestral inference from samples of dna sequences with recombination. *Journal of Computational Biology*, *3*(4), 479–502.

Hein, J., Schierup, M. et Wiuf, C. (2004). *Gene genealogies, variation and evolution : a primer in coalescent theory.* Oxford University Press, USA.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. et Manolio, T. A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, *106*(23), 9362–9367.

Hirschhorn, J. N., Lohmueller, K., Byrne, E. et Hirschhorn, K. (2002). A comprehensive review of genetic association studies. *Genetics in Medicine*, *4*(2), 45–61.

Holmans, P. (2001). Nonparametric linkage. *Handbook of Statistical Genetics*.

Howie, B. N., Donnelly, P. et Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet*, *5*(6), e1000529.

Kingman, J. F. C. (1982). The coalescent. *Stochastic processes and their applications*, *13*(3), 235–248.

Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Annals of eugenics*, *12*(1), 172–175.

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., Team, E. L. P., Christiani, D. C., Wurfel, M. M., Lin, X. *et al.* (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *The American Journal of Human Genetics*, *91*(2), 224–237.

Li, B. et Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases : application to analysis of sequence data. *The American Journal of Human Genetics*, *83*(3), 311–321.

Liu, L., Zhang, D., Liu, H. et Arendt, C. (2013). Robust methods for population stratification in genome wide association studies. *BMC bioinformatics*, *14*(1), 1.

Nordborg, M. et Tavaré, S. (2002). Linkage disequilibrium : what history has to tell us. *TRENDS in Genetics*, *18*(2), 83–90.

Oualkacha, K., Dastani, Z., Li, R., Cingolani, P. E., Spector, T. D., Hammond, C. J., Richards, J. B., Ciampi, A. et Greenwood, C. M. (2013). Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genetic Epidemiology*, *37*(4), 366–376.

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. et Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, *38*(8), 904–909.

Price, A. L., Zaitlen, N. A., Reich, D. et Patterson, N. (2010). New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, *11*(7), 459–463.

Schork, N. J. (1997). Genetics of complex disease : approaches, problems, and solutions. *American journal of respiratory and critical care medicine*, *156*(4), S103–S109.

Schork, N. J., Murray, S. S., Frazer, K. A. et Topol, E. J. (2009). Common vs. rare allele hypotheses for complex diseases. *Current opinion in genetics & development*, *19*(3), 212–219.

Sun, J., Zheng, Y. et Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic epidemiology*, *37*(4), 334–344.

Teare, M. D. et Barrett, J. H. (2005). Genetic linkage studies. *The Lancet*, *366*(9490), 1036–1044.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J. et Lin, X. (2010). Powerful snp-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics*, *86*(6), 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. et Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, *89*(1), 82–93.

Zhang, F., Guo, X. et Deng, H.-W. (2011). Multilocus association testing of quantitative traits based on partial least-squares analysis. *PloS one*, *6*(2), e16739.

Zhang, F., Liu, J., Chen, J. et Deng, H.-W. (2008). Hapsimu : a genetic simulation platform for population-based association studies. *BMC bioinformatics*, *9*(1), 331.

Zhang, S., Zhu, X. et Zhao, H. (2003). On a semiparametric test to detect associations between quantitative traits and candidate genes using unrelated individuals. *Genetic Epidemiology*, *24*(1), 44–56.